Lab Manual: Module 7 – Web Search

Course: Information Retrieval and Web Search
Topics Covered:

- Detecting duplicates and near duplicates using shingling
- Detecting web spam in IR and web search

Submission: Submit both .ipynb file and .ipynb converted to PDF

Submissions with following cases will get a zero

- Code or commented text truncated from the pdf version of the notebook
- Any compilation error in the notebook
- Missing output for any of the programming cells. There should be an output for every code cell

---

Part 1: Detecting Duplicates and Near Duplicates Using Shingling

---

Q1. Document Preparation

Load a set of 4–6 short sample documents/web pages (e.g., product descriptions, blog excerpts).

- Tokenize and normalize the text.
- Preprocess with necessary steps before shingling?
- Use `nltk.word_tokenize()` and lowercase normalization.

In [1]:

```python
# Using wikipedia articles for product descriptions
import wikipedia

titles = ["iPhone 7","Tesla Cybertruck","Tesla Model s",
          "PlayStation 5","iPhone 6", "Playstation 4"]
documents = {}
for title in titles:
    page = wikipedia.page(title)
    documents[title.lower()] = page.summary
documents
```

Out[1]:

```
{'iphone 7': 'The iPhone 4 is a smartphone that was developed and marketed by Apple. It is the fourth generatio
n of the iPhone lineup, succeeding the iPhone 3GS and preceding the iPhone 4s. Following a number of notable le
aks, the iPhone 4 was first unveiled on June 7, 2010, at Apple\'s Worldwide Developers Conference in San Franci
sco, and was released on June 24, 2010, in the United States, United Kingdom, France, Germany, and Japan.\nThe
iPhone 4 introduced a new hardware design to the iPhone family, which Apple\'s CEO Steve Jobs touted as the thi
nnest smartphone in the world at the time; it consisted of a stainless steel frame that doubled as an antenna,
with internal components situated between two panels of aluminosilicate glass. The iPhone 4 introduced Apple\'s
new high-resolution "Retina display" (with a pixel density of 326 pixels per inch) while maintaining the same p
hysical size and aspect ratio as its predecessors, Apple\'s A4 system-on-chip, and with iOS 4—which notably int
roduced multitasking functionality and app folders. It was the first iPhone at the time to include a front-faci
ng camera, which made possible Apple\'s new FaceTime video chat service, and the first to be released in a vers
ion for CDMA networks, ending AT&T\'s period as the exclusive carrier of iPhone products in the United States.\
nThe iPhone 4 received a largely positive reception, with critics praising its revamped design and more powerfu
l hardware, in comparison to previous models. While it was a market success (with over 600,000 pre-orders withi
n 24 hours), the release of the iPhone 4 was plagued by highly publicized reports concerning abnormalities in i
ts new antenna design that caused the device to lose its cellular signal if held in a certain way. Most direct
contact with the phone\'s outer edge would cause a significant decrease in signal strength. Apple released iOS
4.0.1 to try to fix these issues, but were unsuccessful.\nThe iPhone 4 spent the longest time as Apple\'s flags
hip iPhone model at fifteen months. Although the succeeding 4S was announced in October 2011, the 4 continued t
o be sold as a midrange model until September 2012, and thereafter as the entry-level offering in Apple\'s line
up until September 2013 with the announcement of the iPhone 5s and iPhone 5c. The iPhone 4 had one of the longe
st lifespans of any iPhone ever produced, spanning close to four years and available in some developing countri
es until early 2015.',
 'tesla cybertruck': 'The Tesla Cybertruck is a battery-electric full-size pickup truck manufactured by Tesla,
Inc. since 2023. It was first unveiled as a prototype in November 2019, featuring a distinctive angular design
composed of flat, unpainted stainless steel body panels, drawing comparisons to low-polygon computer models.\nO
riginally scheduled for production in late 2021, the vehicle faced multiple delays before entering limited prod
uction at Gigafactory Texas in November 2023, with initial customer deliveries occurring later that month. As o
```

```
f 2025, three variants are available: a tri-motor all-wheel drive (AWD) model marketed as the "Cyberbeast", a d
ual-motor AWD model, and a single-motor rear-wheel drive (RWD) "Long Range" model. EPA range estimates vary by
configuration, from 320 to 350 miles (515 to 565 km). \nAs of 2025, the Cybertruck is sold in the United States
, Mexico, Canada and South Korea. The Cybertruck has been criticized for its production quality and safety conc
erns while its sales have been described as disappointing.',
 'tesla model s': "Tesla, Inc. ( TEZ-lə or   TESS-lə) is an American multinational automotive and clean energy
company. Headquartered in Austin, Texas, it designs, manufactures, and sells battery electric vehicles (BEVs),
stationary battery energy storage devices from home to grid-scale, solar panels and solar shingles, and related
products and services.\nTesla was incorporated in July 2003 by Martin Eberhard and Marc Tarpenning as Tesla Mot
ors. Its name is a tribute to the inventor and electrical engineer Nikola Tesla. In February 2004, Elon Musk le
d Tesla's first funding round and became the company's chairman, subsequently claiming to be a co-founder; in 2
008, he was named chief executive officer. In 2008, the company began production of its first car model, the Ro
adster sports car. This was followed by the Model S sedan in 2012, the Model X SUV in 2015, the Model 3 sedan i
n 2017, the Model Y crossover in 2020, the Tesla Semi truck in 2022, and the Cybertruck pickup truck in 2023.\n
Tesla is one of the world's most valuable companies in terms of market capitalization. Starting in July 2020, i
t has been the world's most valuable automaker. From October 2021 to March 2022, Tesla was a trillion-dollar co
mpany, the seventh US company to reach that valuation. Tesla exceeded $1 trillion in market capitalization agai
n between November 2024 and February 2025. In 2024, the company led the battery electric vehicle market, with 1
7.6% share. In 2023, the company was ranked 69th in the Forbes Global 2000.\nIn May 2025, Tesla again reached a
valuation of $1 trillion, and has been a trillion-dollar company since. In November 2025, Tesla approved a pay
package worth $1 trillion for Musk, which he is to receive over 10 years if he meets specific goals.\nTesla has
been the subject of lawsuits, boycotts, government scrutiny, and journalistic criticism, stemming from allegati
ons of multiple cases of whistleblower retaliation, worker rights violations such as sexual harassment and anti
-union activities, safety defects leading to dozens of recalls, the lack of a public relations department, and
controversial statements from Musk, including overpromising on the company's driving assist technology and prod
uct release timelines.\n\n",
 'playstation 5': 'The PlayStation 2 (PS2) is a home video game console developed and marketed by Sony Computer
Entertainment. It was first released in Japan on 4 March 2000, in North America on October 26, in Europe on Nov
ember 24, in Australia on November 30, and other regions thereafter. It is the successor to the original PlaySt
ation as well as the second installment in the PlayStation brand of consoles. As a sixth-generation console, it
competed with Nintendo\'s GameCube, Sega\'s Dreamcast, and Microsoft\'s Xbox.\nAnnounced in 1999, Sony began de
veloping the console after the immense success of its predecessor. In addition to serving as a game console, it
features a built-in DVD drive and was priced lower than standalone DVD players of the time, enhancing its value
. Full backward compatibility with original PlayStation games and accessories gave it access to a vast launch l
ibrary, far surpassing those of its competitors. The console\'s hardware was also notable for its custom-built
Emotion Engine processor, co-developed with Toshiba, which was promoted as being more powerful than most person
al computers of the era.\nThe PlayStation 2 remains the best-selling video game console of all time, having sol
d 160 million units worldwide, nearly triple the combined sales of competing sixth-generation consoles. It rece
ived widespread critical acclaim and amassed a global library of 10,987 game titles, with 1.54 billion copies s
old. In 2004, Sony revised the console with a smaller, lighter body officially known as the "Slimline". Even af
ter the release of its successor, the PlayStation 3, in 2006, it remained in production and continued to receiv
e new game releases for several years with the last game for the system Pro Evolution Soccer 2014 being release
d in Europe in November 2013. Manufacturing officially ended in early 2013, giving the console one of the longe
st lifespans in video game history.',
 'iphone 6': 'The iPhone 5 is a smartphone that was developed and marketed by Apple. It is the 6th generation i
Phone, succeeding the iPhone 4s, and preceding both the iPhone 5s and iPhone 5c. It was formally unveiled as pa
rt of a press event on September 12, 2012, and subsequently released on September 21, 2012. The iPhone 5 was th
e first iPhone to be announced in September, and setting a trend for subsequent iPhone releases, the first iPho
ne to be completely developed under the guidance of Tim Cook and the last iPhone to be overseen by Steve Jobs.
The iPhone 5\'s design was used three times, first with the iPhone 5 itself in 2012, then with the iPhone 5s in
2013, and finally with the first-generation iPhone SE in 2016.\nThe iPhone 5 featured major design changes in c
omparison to its predecessor. These included an aluminum-based body which was thinner and lighter than previous
models, a taller 4-inch screen with a nearly 16:9 aspect ratio, the Apple A6 system-on-chip, LTE support, and L
ightning, a new compact dock connector which replaced the 30-pin design used by previous iPhone models. This wa
s the second iPhone after the iPhone 4s to include Apple\'s new Sony-made 8 MP camera.\nApple began taking pre-
orders on September 14, 2012, and over two million were received within 24 hours. Initial demand for the iPhone
5 exceeded the supply available at launch on September 21, 2012, and was described by Apple as "extraordinary",
with pre-orders having sold twenty times faster than its predecessors. While reception to the iPhone 5 was gene
rally positive, consumers and reviewers noted hardware issues, such as an unintended purple hue in photos taken
, and the phone\'s coating being prone to chipping. Reception was also mixed over Apple\'s decision to switch t
o a different dock connector design, as the change affected iPhone 5\'s compatibility with accessories that wer
e otherwise compatible with previous iterations of the line.\nAlongside the iPhone 4, the iPhone 5 was official
ly discontinued by Apple on September 10, 2013, with the announcement of its successors, the iPhone 5s and the
iPhone 5c. The iPhone 5 has the joint second-shortest lifespan of any iPhone ever produced with only twelve mon
ths in production, breaking with Apple\'s standard practice of selling an existing iPhone model at a reduced pr
ice upon the release of a new model. This was broken by the iPhone X which only had ten-months in production fr
om November 2017 to September 2018, and tied with the iPhone XS which had twelve-months from September 2018 to
September 2019. The iPhone 11 Pro and subsequent "Pro" designated iPhones have also had twelve month availabili
ty, being discontinued upon release of its successor.\nThe iPhone 5 was replaced as a midrange and then an entr
y-level device by the iPhone 5c; the 5c internal hardware specifications are almost identical to the 5 albeit h
aving a less expensive polycarbonate exterior shell. The iPhone 5 supports iOS 6, 7, 8, 9 and 10. The iPhone 5
does not support iOS 11 due to it dropping support for 32-bit devices. The iPhone 5 is the second iPhone to sup
port five major versions of iOS after the iPhone 4s.',
 'playstation 4': "The PlayStation (codenamed PSX, abbreviated as PS, and retroactively PS1 or PS one) is a hom
e video game console developed and marketed by Sony Computer Entertainment. It was released in Japan on 3 Decem
ber 1994, followed by North America on 9 September 1995, Europe on 29 September 1995, and other regions followi
ng thereafter. As a fifth-generation console, the PlayStation primarily competed with the Nintendo 64 and the S
ega Saturn.\nSony began developing the PlayStation after a failed venture with Nintendo to create a CD-ROM peri
pheral for the Super Nintendo Entertainment System in the early 1990s. The console was primarily designed by Ke
n Kutaragi and Sony Computer Entertainment in Japan, while additional development was outsourced in the United
Kingdom. An emphasis on 3D polygon graphics was placed at the forefront of the console's design. PlayStation ga
me production was designed to be streamlined and inclusive, enticing the support of many third party developers
.\nThe console proved popular for its extensive game library, popular franchises, low retail price, and aggress
```

```
ive youth marketing which advertised it as the preferable console for adolescents and adults. Critically acclai
med games that defined the console include Gran Turismo, Crash Bandicoot, Spyro the Dragon, Tomb Raider, Reside
nt Evil, Metal Gear Solid, Tekken 3, and Final Fantasy VII. Sony ceased production of the PlayStation on 23 Mar
ch 2006—over eleven years after it had been released, and in the same year the PlayStation 3 debuted. More than
4,000 PlayStation games were released, with cumulative sales of 962 million units.\nThe PlayStation signaled So
ny's rise to power in the video game industry. It received acclaim and sold strongly; in less than a decade, it
became the first computer entertainment platform to ship over 100 million units. Its use of compact discs heral
ded the game industry's transition from cartridges. The PlayStation's success led to a line of successors, begi
nning with the PlayStation 2 in 2000. In the same year, Sony released a smaller and cheaper model, the PS one."
}
```

In [2]:

```python
# TOKENIZING AND NORMALIZING DOCUMENTS
import nltk
tokenized_normalized = {}
for title, summary in documents.items():
    tokenized_doc = nltk.word_tokenize(summary)
    lowercased_doc = [token.lower() for token in tokenized_doc]
    normalized_doc = [
        token for token in lowercased_doc if token.isalpha() and token.isascii()
    ]
    tokenized_normalized[title] = normalized_doc
print(tokenized_normalized)
```

```
{'iphone 7': ['the', 'iphone', 'is', 'a', 'smartphone', 'that', 'was', 'developed', 'and', 'marketed', 'by', 'ap
ple', 'it', 'is', 'the', 'fourth', 'generation', 'of', 'the', 'iphone', 'lineup', 'succeeding', 'the', 'iphone',
'and', 'preceding', 'the', 'iphone', 'following', 'a', 'number', 'of', 'notable', 'leaks', 'the', 'iphone', 'was
', 'first', 'unveiled', 'on', 'june', 'at', 'apple', 'worldwide', 'developers', 'conference', 'in', 'san', 'fran
cisco', 'and', 'was', 'released', 'on', 'june', 'in', 'the', 'united', 'states', 'united', 'kingdom', 'france',
'germany', 'and', 'japan', 'the', 'iphone', 'introduced', 'a', 'new', 'hardware', 'design', 'to', 'the', 'iphone
', 'family', 'which', 'apple', 'ceo', 'steve', 'jobs', 'touted', 'as', 'the', 'thinnest', 'smartphone', 'in', 't
he', 'world', 'at', 'the', 'time', 'it', 'consisted', 'of', 'a', 'stainless', 'steel', 'frame', 'that', 'doubled
', 'as', 'an', 'antenna', 'with', 'internal', 'components', 'situated', 'between', 'two', 'panels', 'of', 'alumi
nosilicate', 'glass', 'the', 'iphone', 'introduced', 'apple', 'new', 'retina', 'display', 'with', 'a', 'pixel',
'density', 'of', 'pixels', 'per', 'inch', 'while', 'maintaining', 'the', 'same', 'physical', 'size', 'and', 'asp
ect', 'ratio', 'as', 'its', 'predecessors', 'apple', 'and', 'with', 'ios', 'notably', 'introduced', 'multitaskin
g', 'functionality', 'and', 'app', 'folders', 'it', 'was', 'the', 'first', 'iphone', 'at', 'the', 'time', 'to',
'include', 'a', 'camera', 'which', 'made', 'possible', 'apple', 'new', 'facetime', 'video', 'chat', 'service', '
and', 'the', 'first', 'to', 'be', 'released', 'in', 'a', 'version', 'for', 'cdma', 'networks', 'ending', 'at', '
t', 'period', 'as', 'the', 'exclusive', 'carrier', 'of', 'iphone', 'products', 'in', 'the', 'united', 'states',
'the', 'iphone', 'received', 'a', 'largely', 'positive', 'reception', 'with', 'critics', 'praising', 'its', 'rev
amped', 'design', 'and', 'more', 'powerful', 'hardware', 'in', 'comparison', 'to', 'previous', 'models', 'while'
, 'it', 'was', 'a', 'market', 'success', 'with', 'over', 'within', 'hours', 'the', 'release', 'of', 'the', 'ipho
ne', 'was', 'plagued', 'by', 'highly', 'publicized', 'reports', 'concerning', 'abnormalities', 'in', 'its', 'new
', 'antenna', 'design', 'that', 'caused', 'the', 'device', 'to', 'lose', 'its', 'cellular', 'signal', 'if', 'hel
d', 'in', 'a', 'certain', 'way', 'most', 'direct', 'contact', 'with', 'the', 'phone', 'outer', 'edge', 'would',
'cause', 'a', 'significant', 'decrease', 'in', 'signal', 'strength', 'apple', 'released', 'ios', 'to', 'try', 't
o', 'fix', 'these', 'issues', 'but', 'were', 'unsuccessful', 'the', 'iphone', 'spent', 'the', 'longest', 'time',
'as', 'apple', 'flagship', 'iphone', 'model', 'at', 'fifteen', 'months', 'although', 'the', 'succeeding', 'was',
'announced', 'in', 'october', 'the', 'continued', 'to', 'be', 'sold', 'as', 'a', 'midrange', 'model', 'until', '
september', 'and', 'thereafter', 'as', 'the', 'offering', 'in', 'apple', 'lineup', 'until', 'september', 'with',
'the', 'announcement', 'of', 'the', 'iphone', 'and', 'iphone', 'the', 'iphone', 'had', 'one', 'of', 'the', 'long
est', 'lifespans', 'of', 'any', 'iphone', 'ever', 'produced', 'spanning', 'close', 'to', 'four', 'years', 'and',
'available', 'in', 'some', 'developing', 'countries', 'until', 'early'], 'tesla cybertruck': ['the', 'tesla', 'c
ybertruck', 'is', 'a', 'pickup', 'truck', 'manufactured', 'by', 'tesla', 'since', 'it', 'was', 'first', 'unveile
d', 'as', 'a', 'prototype', 'in', 'november', 'featuring', 'a', 'distinctive', 'angular', 'design', 'composed',
'of', 'flat', 'unpainted', 'stainless', 'steel', 'body', 'panels', 'drawing', 'comparisons', 'to', 'computer', '
models', 'originally', 'scheduled', 'for', 'production', 'in', 'late', 'the', 'vehicle', 'faced', 'multiple', 'd
elays', 'before', 'entering', 'limited', 'production', 'at', 'gigafactory', 'texas', 'in', 'november', 'with', '
initial', 'customer', 'deliveries', 'occurring', 'later', 'that', 'month', 'as', 'of', 'three', 'variants', 'are
', 'available', 'a', 'drive', 'awd', 'model', 'marketed', 'as', 'the', 'cyberbeast', 'a', 'awd', 'model', 'and',
'a', 'drive', 'rwd', 'long', 'range', 'model', 'epa', 'range', 'estimates', 'vary', 'by', 'configuration', 'from
', 'to', 'miles', 'to', 'km', 'as', 'of', 'the', 'cybertruck', 'is', 'sold', 'in', 'the', 'united', 'states', 'm
exico', 'canada', 'and', 'south', 'korea', 'the', 'cybertruck', 'has', 'been', 'criticized', 'for', 'its', 'prod
uction', 'quality', 'and', 'safety', 'concerns', 'while', 'its', 'sales', 'have', 'been', 'described', 'as', 'di
sappointing'], 'tesla model s': ['tesla', 'or', 'is', 'an', 'american', 'multinational', 'automotive', 'and', 'c
lean', 'energy', 'company', 'headquartered', 'in', 'austin', 'texas', 'it', 'designs', 'manufactures', 'and', 's
ells', 'battery', 'electric', 'vehicles', 'bevs', 'stationary', 'battery', 'energy', 'storage', 'devices', 'from
', 'home', 'to', 'solar', 'panels', 'and', 'solar', 'shingles', 'and', 'related', 'products', 'and', 'services',
'tesla', 'was', 'incorporated', 'in', 'july', 'by', 'martin', 'eberhard', 'and', 'marc', 'tarpenning', 'as', 'te
sla', 'motors', 'its', 'name', 'is', 'a', 'tribute', 'to', 'the', 'inventor', 'and', 'electrical', 'engineer', '
nikola', 'tesla', 'in', 'february', 'elon', 'musk', 'led', 'tesla', 'first', 'funding', 'round', 'and', 'became'
, 'the', 'company', 'chairman', 'subsequently', 'claiming', 'to', 'be', 'a', 'in', 'he', 'was', 'named', 'chief'
, 'executive', 'officer', 'in', 'the', 'company', 'began', 'production', 'of', 'its', 'first', 'car', 'model', '
the', 'roadster', 'sports', 'car', 'this', 'was', 'followed', 'by', 'the', 'model', 's', 'sedan', 'in', 'the', '
model', 'x', 'suv', 'in', 'the', 'model', 'sedan', 'in', 'the', 'model', 'y', 'crossover', 'in', 'the', 'tesla',
'semi', 'truck', 'in', 'and', 'the', 'cybertruck', 'pickup', 'truck', 'in', 'tesla', 'is', 'one', 'of', 'the', '
world', 'most', 'valuable', 'companies', 'in', 'terms', 'of', 'market', 'capitalization', 'starting', 'in', 'jul
y', 'it', 'has', 'been', 'the', 'world', 'most', 'valuable', 'automaker', 'from', 'october', 'to', 'march', 'tes
la', 'was', 'a', 'company', 'the', 'seventh', 'us', 'company', 'to', 'reach', 'that', 'valuation', 'tesla', 'exc
eeded', 'trillion', 'in', 'market', 'capitalization', 'again', 'between', 'november', 'and', 'february', 'in', '
the', 'company', 'led', 'the', 'battery', 'electric', 'vehicle', 'market', 'with', 'share', 'in', 'the', 'compan
y', 'was', 'ranked', 'in', 'the', 'forbes', 'global', 'in', 'may', 'tesla', 'again', 'reached', 'a', 'valuation'
, 'of', 'trillion', 'and', 'has', 'been', 'a', 'company', 'since', 'in', 'november', 'tesla', 'approved', 'a', '
pay', 'package', 'worth', 'trillion', 'for', 'musk', 'which', 'he', 'is', 'to', 'receive', 'over', 'years', 'if'
```

```
, 'he', 'meets', 'specific', 'goals', 'tesla', 'has', 'been', 'the', 'subject', 'of', 'lawsuits', 'boycotts', 'g
overnment', 'scrutiny', 'and', 'journalistic', 'criticism', 'stemming', 'from', 'allegations', 'of', 'multiple',
'cases', 'of', 'whistleblower', 'retaliation', 'worker', 'rights', 'violations', 'such', 'as', 'sexual', 'harass
ment', 'and', 'activities', 'safety', 'defects', 'leading', 'to', 'dozens', 'of', 'recalls', 'the', 'lack', 'of'
, 'a', 'public', 'relations', 'department', 'and', 'controversial', 'statements', 'from', 'musk', 'including', '
overpromising', 'on', 'the', 'company', 'driving', 'assist', 'technology', 'and', 'product', 'release', 'timelin
es'], 'playstation 5': ['the', 'playstation', 'is', 'a', 'home', 'video', 'game', 'console', 'developed', 'and',
'marketed', 'by', 'sony', 'computer', 'entertainment', 'it', 'was', 'first', 'released', 'in', 'japan', 'on', 'm
arch', 'in', 'north', 'america', 'on', 'october', 'in', 'europe', 'on', 'november', 'in', 'australia', 'on', 'no
vember', 'and', 'other', 'regions', 'thereafter', 'it', 'is', 'the', 'successor', 'to', 'the', 'original', 'play
station', 'as', 'well', 'as', 'the', 'second', 'installment', 'in', 'the', 'playstation', 'brand', 'of', 'consol
es', 'as', 'a', 'console', 'it', 'competed', 'with', 'nintendo', 'gamecube', 'sega', 'dreamcast', 'and', 'micros
oft', 'xbox', 'announced', 'in', 'sony', 'began', 'developing', 'the', 'console', 'after', 'the', 'immense', 'su
ccess', 'of', 'its', 'predecessor', 'in', 'addition', 'to', 'serving', 'as', 'a', 'game', 'console', 'it', 'feat
ures', 'a', 'dvd', 'drive', 'and', 'was', 'priced', 'lower', 'than', 'standalone', 'dvd', 'players', 'of', 'the'
, 'time', 'enhancing', 'its', 'value', 'full', 'backward', 'compatibility', 'with', 'original', 'playstation', '
games', 'and', 'accessories', 'gave', 'it', 'access', 'to', 'a', 'vast', 'launch', 'library', 'far', 'surpassing
', 'those', 'of', 'its', 'competitors', 'the', 'console', 'hardware', 'was', 'also', 'notable', 'for', 'its', 'e
motion', 'engine', 'processor', 'with', 'toshiba', 'which', 'was', 'promoted', 'as', 'being', 'more', 'powerful'
, 'than', 'most', 'personal', 'computers', 'of', 'the', 'era', 'the', 'playstation', 'remains', 'the', 'video',
'game', 'console', 'of', 'all', 'time', 'having', 'sold', 'million', 'units', 'worldwide', 'nearly', 'triple', '
the', 'combined', 'sales', 'of', 'competing', 'consoles', 'it', 'received', 'widespread', 'critical', 'acclaim',
'and', 'amassed', 'a', 'global', 'library', 'of', 'game', 'titles', 'with', 'billion', 'copies', 'sold', 'in', '
sony', 'revised', 'the', 'console', 'with', 'a', 'smaller', 'lighter', 'body', 'officially', 'known', 'as', 'the
', 'slimline', 'even', 'after', 'the', 'release', 'of', 'its', 'successor', 'the', 'playstation', 'in', 'it', 'r
emained', 'in', 'production', 'and', 'continued', 'to', 'receive', 'new', 'game', 'releases', 'for', 'several',
'years', 'with', 'the', 'last', 'game', 'for', 'the', 'system', 'pro', 'evolution', 'soccer', 'being', 'released
', 'in', 'europe', 'in', 'november', 'manufacturing', 'officially', 'ended', 'in', 'early', 'giving', 'the', 'co
nsole', 'one', 'of', 'the', 'longest', 'lifespans', 'in', 'video', 'game', 'history'], 'iphone 6': ['the', 'ipho
ne', 'is', 'a', 'smartphone', 'that', 'was', 'developed', 'and', 'marketed', 'by', 'apple', 'it', 'is', 'the', '
generation', 'iphone', 'succeeding', 'the', 'iphone', 'and', 'preceding', 'both', 'the', 'iphone', 'and', 'iphon
e', 'it', 'was', 'formally', 'unveiled', 'as', 'part', 'of', 'a', 'press', 'event', 'on', 'september', 'and', 's
ubsequently', 'released', 'on', 'september', 'the', 'iphone', 'was', 'the', 'first', 'iphone', 'to', 'be', 'anno
unced', 'in', 'september', 'and', 'setting', 'a', 'trend', 'for', 'subsequent', 'iphone', 'releases', 'the', 'fi
rst', 'iphone', 'to', 'be', 'completely', 'developed', 'under', 'the', 'guidance', 'of', 'tim', 'cook', 'and', '
the', 'last', 'iphone', 'to', 'be', 'overseen', 'by', 'steve', 'jobs', 'the', 'iphone', 'design', 'was', 'used',
'three', 'times', 'first', 'with', 'the', 'iphone', 'itself', 'in', 'then', 'with', 'the', 'iphone', 'in', 'and'
, 'finally', 'with', 'the', 'iphone', 'se', 'in', 'the', 'iphone', 'featured', 'major', 'design', 'changes', 'in
', 'comparison', 'to', 'its', 'predecessor', 'these', 'included', 'an', 'body', 'which', 'was', 'thinner', 'and'
, 'lighter', 'than', 'previous', 'models', 'a', 'taller', 'screen', 'with', 'a', 'nearly', 'aspect', 'ratio', 't
he', 'apple', 'lte', 'support', 'and', 'lightning', 'a', 'new', 'compact', 'dock', 'connector', 'which', 'replac
ed', 'the', 'design', 'used', 'by', 'previous', 'iphone', 'models', 'this', 'was', 'the', 'second', 'iphone', 'a
fter', 'the', 'iphone', 'to', 'include', 'apple', 'new', 'mp', 'camera', 'apple', 'began', 'taking', 'on', 'sept
ember', 'and', 'over', 'two', 'million', 'were', 'received', 'within', 'hours', 'initial', 'demand', 'for', 'the
', 'iphone', 'exceeded', 'the', 'supply', 'available', 'at', 'launch', 'on', 'september', 'and', 'was', 'describ
ed', 'by', 'apple', 'as', 'extraordinary', 'with', 'having', 'sold', 'twenty', 'times', 'faster', 'than', 'its',
'predecessors', 'while', 'reception', 'to', 'the', 'iphone', 'was', 'generally', 'positive', 'consumers', 'and',
'reviewers', 'noted', 'hardware', 'issues', 'such', 'as', 'an', 'unintended', 'purple', 'hue', 'in', 'photos', '
taken', 'and', 'the', 'phone', 'coating', 'being', 'prone', 'to', 'chipping', 'reception', 'was', 'also', 'mixed
', 'over', 'apple', 'decision', 'to', 'switch', 'to', 'a', 'different', 'dock', 'connector', 'design', 'as', 'th
e', 'change', 'affected', 'iphone', 'compatibility', 'with', 'accessories', 'that', 'were', 'otherwise', 'compat
ible', 'with', 'previous', 'iterations', 'of', 'the', 'line', 'alongside', 'the', 'iphone', 'the', 'iphone', 'wa
s', 'officially', 'discontinued', 'by', 'apple', 'on', 'september', 'with', 'the', 'announcement', 'of', 'its',
'successors', 'the', 'iphone', 'and', 'the', 'iphone', 'the', 'iphone', 'has', 'the', 'joint', 'lifespan', 'of',
'any', 'iphone', 'ever', 'produced', 'with', 'only', 'twelve', 'months', 'in', 'production', 'breaking', 'with',
'apple', 'standard', 'practice', 'of', 'selling', 'an', 'existing', 'iphone', 'model', 'at', 'a', 'reduced', 'pr
ice', 'upon', 'the', 'release', 'of', 'a', 'new', 'model', 'this', 'was', 'broken', 'by', 'the', 'iphone', 'x',
'which', 'only', 'had', 'in', 'production', 'from', 'november', 'to', 'september', 'and', 'tied', 'with', 'the',
'iphone', 'xs', 'which', 'had', 'from', 'september', 'to', 'september', 'the', 'iphone', 'pro', 'and', 'subseque
nt', 'pro', 'designated', 'iphones', 'have', 'also', 'had', 'twelve', 'month', 'availability', 'being', 'discont
inued', 'upon', 'release', 'of', 'its', 'successor', 'the', 'iphone', 'was', 'replaced', 'as', 'a', 'midrange',
'and', 'then', 'an', 'device', 'by', 'the', 'iphone', 'the', 'internal', 'hardware', 'specifications', 'are', 'a
lmost', 'identical', 'to', 'the', 'albeit', 'having', 'a', 'less', 'expensive', 'polycarbonate', 'exterior', 'sh
ell', 'the', 'iphone', 'supports', 'ios', 'and', 'the', 'iphone', 'does', 'not', 'support', 'ios', 'due', 'to',
'it', 'dropping', 'support', 'for', 'devices', 'the', 'iphone', 'is', 'the', 'second', 'iphone', 'to', 'support'
, 'five', 'major', 'versions', 'of', 'ios', 'after', 'the', 'iphone'], 'playstation 4': ['the', 'playstation', '
codenamed', 'psx', 'abbreviated', 'as', 'ps', 'and', 'retroactively', 'or', 'ps', 'one', 'is', 'a', 'home', 'vid
eo', 'game', 'console', 'developed', 'and', 'marketed', 'by', 'sony', 'computer', 'entertainment', 'it', 'was',
'released', 'in', 'japan', 'on', 'december', 'followed', 'by', 'north', 'america', 'on', 'september', 'europe',
'on', 'september', 'and', 'other', 'regions', 'following', 'thereafter', 'as', 'a', 'console', 'the', 'playstati
on', 'primarily', 'competed', 'with', 'the', 'nintendo', 'and', 'the', 'sega', 'saturn', 'sony', 'began', 'devel
oping', 'the', 'playstation', 'after', 'a', 'failed', 'venture', 'with', 'nintendo', 'to', 'create', 'a', 'perip
heral', 'for', 'the', 'super', 'nintendo', 'entertainment', 'system', 'in', 'the', 'early', 'the', 'console', 'w
as', 'primarily', 'designed', 'by', 'ken', 'kutaragi', 'and', 'sony', 'computer', 'entertainment', 'in', 'japan'
, 'while', 'additional', 'development', 'was', 'outsourced', 'in', 'the', 'united', 'kingdom', 'an', 'emphasis',
'on', 'polygon', 'graphics', 'was', 'placed', 'at', 'the', 'forefront', 'of', 'the', 'console', 'design', 'plays
tation', 'game', 'production', 'was', 'designed', 'to', 'be', 'streamlined', 'and', 'inclusive', 'enticing', 'th
e', 'support', 'of', 'many', 'third', 'party', 'developers', 'the', 'console', 'proved', 'popular', 'for', 'its'
, 'extensive', 'game', 'library', 'popular', 'franchises', 'low', 'retail', 'price', 'and', 'aggressive', 'youth
', 'marketing', 'which', 'advertised', 'it', 'as', 'the', 'preferable', 'console', 'for', 'adolescents', 'and',
'adults', 'critically', 'acclaimed', 'games', 'that', 'defined', 'the', 'console', 'include', 'gran', 'turismo',
'crash', 'bandicoot', 'spyro', 'the', 'dragon', 'tomb', 'raider', 'resident', 'evil', 'metal', 'gear', 'solid',
'tekken', 'and', 'final', 'fantasy', 'vii', 'sony', 'ceased', 'production', 'of', 'the', 'playstation', 'on', 'm
```

```
arch', 'eleven', 'years', 'after', 'it', 'had', 'been', 'released', 'and', 'in', 'the', 'same', 'year', 'the', '
playstation', 'debuted', 'more', 'than', 'playstation', 'games', 'were', 'released', 'with', 'cumulative', 'sale
s', 'of', 'million', 'units', 'the', 'playstation', 'signaled', 'sony', 'rise', 'to', 'power', 'in', 'the', 'vid
eo', 'game', 'industry', 'it', 'received', 'acclaim', 'and', 'sold', 'strongly', 'in', 'less', 'than', 'a', 'dec
ade', 'it', 'became', 'the', 'first', 'computer', 'entertainment', 'platform', 'to', 'ship', 'over', 'million',
'units', 'its', 'use', 'of', 'compact', 'discs', 'heralded', 'the', 'game', 'industry', 'transition', 'from', 'c
artridges', 'the', 'playstation', 'success', 'led', 'to', 'a', 'line', 'of', 'successors', 'beginning', 'with',
'the', 'playstation', 'in', 'in', 'the', 'same', 'year', 'sony', 'released', 'a', 'smaller', 'and', 'cheaper', '
model', 'the', 'ps', 'one']}
```

Q2. Shingle Generation

Generate k-shingle sets.

- Choose a value for k (e.g., 3 or 4).
- Generate shingles for each document using consecutive sequences of k terms.
- Explain how the choice of k affects sensitivity to small changes.

In [3]:

```python
# Shingle fuction, default as k=3
def shingle(document_tokens, k=3):
    shingles = []
    for idx in range(len(document_tokens) - k +1):
        shingle = tuple(document_tokens[idx:idx+k])
        shingles.append(shingle)
    return shingles
```

In [4]:

```python
# shingle every document and save in dict
shingled_docs = {}
for title, doc in tokenized_normalized.items():
    shingled_doc = shingle(doc)
    shingled_docs[title] = shingled_doc
print(shingled_docs)
```

```
{'iphone 7': [('the', 'iphone', 'is'), ('iphone', 'is', 'a'), ('is', 'a', 'smartphone'), ('a', 'smartphone', 'th
at'), ('smartphone', 'that', 'was'), ('that', 'was', 'developed'), ('was', 'developed', 'and'), ('developed', 'a
nd', 'marketed'), ('and', 'marketed', 'by'), ('marketed', 'by', 'apple'), ('by', 'apple', 'it'), ('apple', 'it',
'is'), ('it', 'is', 'the'), ('is', 'the', 'fourth'), ('the', 'fourth', 'generation'), ('fourth', 'generation', '
of'), ('generation', 'of', 'the'), ('of', 'the', 'iphone'), ('the', 'iphone', 'lineup'), ('iphone', 'lineup', 's
ucceeding'), ('lineup', 'succeeding', 'the'), ('succeeding', 'the', 'iphone'), ('the', 'iphone', 'and'), ('iphon
e', 'and', 'preceding'), ('and', 'preceding', 'the'), ('preceding', 'the', 'iphone'), ('the', 'iphone', 'followi
ng'), ('iphone', 'following', 'a'), ('following', 'a', 'number'), ('a', 'number', 'of'), ('number', 'of', 'notab
le'), ('of', 'notable', 'leaks'), ('notable', 'leaks', 'the'), ('leaks', 'the', 'iphone'), ('the', 'iphone', 'wa
s'), ('iphone', 'was', 'first'), ('was', 'first', 'unveiled'), ('first', 'unveiled', 'on'), ('unveiled', 'on', '
june'), ('on', 'june', 'at'), ('june', 'at', 'apple'), ('at', 'apple', 'worldwide'), ('apple', 'worldwide', 'dev
elopers'), ('worldwide', 'developers', 'conference'), ('developers', 'conference', 'in'), ('conference', 'in', '
san'), ('in', 'san', 'francisco'), ('san', 'francisco', 'and'), ('francisco', 'and', 'was'), ('and', 'was', 'rel
eased'), ('was', 'released', 'on'), ('released', 'on', 'june'), ('on', 'june', 'in'), ('june', 'in', 'the'), ('i
n', 'the', 'united'), ('the', 'united', 'states'), ('united', 'states', 'united'), ('states', 'united', 'kingdom
'), ('united', 'kingdom', 'france'), ('kingdom', 'france', 'germany'), ('france', 'germany', 'and'), ('germany',
'and', 'japan'), ('and', 'japan', 'the'), ('japan', 'the', 'iphone'), ('the', 'iphone', 'introduced'), ('iphone'
, 'introduced', 'a'), ('introduced', 'a', 'new'), ('a', 'new', 'hardware'), ('new', 'hardware', 'design'), ('har
dware', 'design', 'to'), ('design', 'to', 'the'), ('to', 'the', 'iphone'), ('the', 'iphone', 'family'), ('iphone
', 'family', 'which'), ('family', 'which', 'apple'), ('which', 'apple', 'ceo'), ('apple', 'ceo', 'steve'), ('ceo
', 'steve', 'jobs'), ('steve', 'jobs', 'touted'), ('jobs', 'touted', 'as'), ('touted', 'as', 'the'), ('as', 'the
', 'thinnest'), ('the', 'thinnest', 'smartphone'), ('thinnest', 'smartphone', 'in'), ('smartphone', 'in', 'the')
, ('in', 'the', 'world'), ('the', 'world', 'at'), ('world', 'at', 'the'), ('at', 'the', 'time'), ('the', 'time',
'it'), ('time', 'it', 'consisted'), ('it', 'consisted', 'of'), ('consisted', 'of', 'a'), ('of', 'a', 'stainless'
), ('a', 'stainless', 'steel'), ('stainless', 'steel', 'frame'), ('steel', 'frame', 'that'), ('frame', 'that', '
doubled'), ('that', 'doubled', 'as'), ('doubled', 'as', 'an'), ('as', 'an', 'antenna'), ('an', 'antenna', 'with'
), ('antenna', 'with', 'internal'), ('with', 'internal', 'components'), ('internal', 'components', 'situated'),
('components', 'situated', 'between'), ('situated', 'between', 'two'), ('between', 'two', 'panels'), ('two', 'pa
nels', 'of'), ('panels', 'of', 'aluminosilicate'), ('of', 'aluminosilicate', 'glass'), ('aluminosilicate', 'glas
s', 'the'), ('glass', 'the', 'iphone'), ('the', 'iphone', 'introduced'), ('iphone', 'introduced', 'apple'), ('in
troduced', 'apple', 'new'), ('apple', 'new', 'retina'), ('new', 'retina', 'display'), ('retina', 'display', 'wit
h'), ('display', 'with', 'a'), ('with', 'a', 'pixel'), ('a', 'pixel', 'density'), ('pixel', 'density', 'of'), ('
density', 'of', 'pixels'), ('of', 'pixels', 'per'), ('pixels', 'per', 'inch'), ('per', 'inch', 'while'), ('inch'
, 'while', 'maintaining'), ('while', 'maintaining', 'the'), ('maintaining', 'the', 'same'), ('the', 'same', 'phy
sical'), ('same', 'physical', 'size'), ('physical', 'size', 'and'), ('size', 'and', 'aspect'), ('and', 'aspect',
'ratio'), ('aspect', 'ratio', 'as'), ('ratio', 'as', 'its'), ('as', 'its', 'predecessors'), ('its', 'predecessor
s', 'apple'), ('predecessors', 'apple', 'and'), ('apple', 'and', 'with'), ('and', 'with', 'ios'), ('with', 'ios'
, 'notably'), ('ios', 'notably', 'introduced'), ('notably', 'introduced', 'multitasking'), ('introduced', 'multi
tasking', 'functionality'), ('multitasking', 'functionality', 'and'), ('functionality', 'and', 'app'), ('and', '
app', 'folders'), ('app', 'folders', 'it'), ('folders', 'it', 'was'), ('it', 'was', 'the'), ('was', 'the', 'firs
t'), ('the', 'first', 'iphone'), ('first', 'iphone', 'at'), ('iphone', 'at', 'the'), ('at', 'the', 'time'), ('th
e', 'time', 'to'), ('time', 'to', 'include'), ('to', 'include', 'a'), ('include', 'a', 'camera'), ('a', 'camera'
, 'which'), ('camera', 'which', 'made'), ('which', 'made', 'possible'), ('made', 'possible', 'apple'), ('possibl
e', 'apple', 'new'), ('apple', 'new', 'facetime'), ('new', 'facetime', 'video'), ('facetime', 'video', 'chat'),
('video', 'chat', 'service'), ('chat', 'service', 'and'), ('service', 'and', 'the'), ('and', 'the', 'first'), ('
the', 'first', 'to'), ('first', 'to', 'be'), ('to', 'be', 'released'), ('be', 'released', 'in'), ('released', 'i
n', 'a'), ('in', 'a', 'version'), ('a', 'version', 'for'), ('version', 'for', 'cdma'), ('for', 'cdma', 'networks
'), ('cdma', 'networks', 'ending'), ('networks', 'ending', 'at'), ('ending', 'at', 't'), ('at', 't', 'period'),
('t', 'period', 'as'), ('period', 'as', 'the'), ('as', 'the', 'exclusive'), ('the', 'exclusive', 'carrier'), ('e
xclusive', 'carrier', 'of'), ('carrier', 'of', 'iphone'), ('of', 'iphone', 'products'), ('iphone', 'products', '
in'), ('products', 'in', 'the'), ('in', 'the', 'united'), ('the', 'united', 'states'), ('united', 'states', 'the
'), ('states', 'the', 'iphone'), ('the', 'iphone', 'received'), ('iphone', 'received', 'a'), ('received', 'a', '
largely'), ('a', 'largely', 'positive'), ('largely', 'positive', 'reception'), ('positive', 'reception', 'with')
, ('reception', 'with', 'critics'), ('with', 'critics', 'praising'), ('critics', 'praising', 'its'), ('praising'
, 'its', 'revamped'), ('its', 'revamped', 'design'), ('revamped', 'design', 'and'), ('design', 'and', 'more'), (
'and', 'more', 'powerful'), ('more', 'powerful', 'hardware'), ('powerful', 'hardware', 'in'), ('hardware', 'in',
'comparison'), ('in', 'comparison', 'to'), ('comparison', 'to', 'previous'), ('to', 'previous', 'models'), ('pre
vious', 'models', 'while'), ('models', 'while', 'it'), ('while', 'it', 'was'), ('it', 'was', 'a'), ('was', 'a',
'market'), ('a', 'market', 'success'), ('market', 'success', 'with'), ('success', 'with', 'over'), ('with', 'ove
r', 'within'), ('over', 'within', 'hours'), ('within', 'hours', 'the'), ('hours', 'the', 'release'), ('the', 're
lease', 'of'), ('release', 'of', 'the'), ('of', 'the', 'iphone'), ('the', 'iphone', 'was'), ('iphone', 'was', 'p
lagued'), ('was', 'plagued', 'by'), ('plagued', 'by', 'highly'), ('by', 'highly', 'publicized'), ('highly', 'pub
licized', 'reports'), ('publicized', 'reports', 'concerning'), ('reports', 'concerning', 'abnormalities'), ('con
cerning', 'abnormalities', 'in'), ('abnormalities', 'in', 'its'), ('in', 'its', 'new'), ('its', 'new', 'antenna'
), ('new', 'antenna', 'design'), ('antenna', 'design', 'that'), ('design', 'that', 'caused'), ('that', 'caused',
'the'), ('caused', 'the', 'device'), ('the', 'device', 'to'), ('device', 'to', 'lose'), ('to', 'lose', 'its'), (
'lose', 'its', 'cellular'), ('its', 'cellular', 'signal'), ('cellular', 'signal', 'if'), ('signal', 'if', 'held'
), ('if', 'held', 'in'), ('held', 'in', 'a'), ('in', 'a', 'certain'), ('a', 'certain', 'way'), ('certain', 'way'
, 'most'), ('way', 'most', 'direct'), ('most', 'direct', 'contact'), ('direct', 'contact', 'with'), ('contact',
'with', 'the'), ('with', 'the', 'phone'), ('the', 'phone', 'outer'), ('phone', 'outer', 'edge'), ('outer', 'edge
', 'would'), ('edge', 'would', 'cause'), ('would', 'cause', 'a'), ('cause', 'a', 'significant'), ('a', 'signific
ant', 'decrease'), ('significant', 'decrease', 'in'), ('decrease', 'in', 'signal'), ('in', 'signal', 'strength')
, ('signal', 'strength', 'apple'), ('strength', 'apple', 'released'), ('apple', 'released', 'ios'), ('released',
'ios', 'to'), ('ios', 'to', 'try'), ('to', 'try', 'to'), ('try', 'to', 'fix'), ('to', 'fix', 'these'), ('fix', '
these', 'issues'), ('these', 'issues', 'but'), ('issues', 'but', 'were'), ('but', 'were', 'unsuccessful'), ('wer
e', 'unsuccessful', 'the'), ('unsuccessful', 'the', 'iphone'), ('the', 'iphone', 'spent'), ('iphone', 'spent', '
the'), ('spent', 'the', 'longest'), ('the', 'longest', 'time'), ('longest', 'time', 'as'), ('time', 'as', 'apple
'), ('as', 'apple', 'flagship'), ('apple', 'flagship', 'iphone'), ('flagship', 'iphone', 'model'), ('iphone', 'm
odel', 'at'), ('model', 'at', 'fifteen'), ('at', 'fifteen', 'months'), ('fifteen', 'months', 'although'), ('mont
hs', 'although', 'the'), ('although', 'the', 'succeeding'), ('the', 'succeeding', 'was'), ('succeeding', 'was',
'announced'), ('was', 'announced', 'in'), ('announced', 'in', 'october'), ('in', 'october', 'the'), ('october',
'the', 'continued'), ('the', 'continued', 'to'), ('continued', 'to', 'be'), ('to', 'be', 'sold'), ('be', 'sold',
'as'), ('sold', 'as', 'a'), ('as', 'a', 'midrange'), ('a', 'midrange', 'model'), ('midrange', 'model', 'until'),
('model', 'until', 'september'), ('until', 'september', 'and'), ('september', 'and', 'thereafter'), ('and', 'the
reafter', 'as'), ('thereafter', 'as', 'the'), ('as', 'the', 'offering'), ('the', 'offering', 'in'), ('offering',
'in', 'apple'), ('in', 'apple', 'lineup'), ('apple', 'lineup', 'until'), ('lineup', 'until', 'september'), ('unt
il', 'september', 'with'), ('september', 'with', 'the'), ('with', 'the', 'announcement'), ('the', 'announcement'
, 'of'), ('announcement', 'of', 'the'), ('of', 'the', 'iphone'), ('the', 'iphone', 'and'), ('iphone', 'and', 'ip
hone'), ('and', 'iphone', 'the'), ('iphone', 'the', 'iphone'), ('the', 'iphone', 'had'), ('iphone', 'had', 'one'
), ('had', 'one', 'of'), ('one', 'of', 'the'), ('of', 'the', 'longest'), ('the', 'longest', 'lifespans'), ('long
est', 'lifespans', 'of'), ('lifespans', 'of', 'any'), ('of', 'any', 'iphone'), ('any', 'iphone', 'ever'), ('ipho
ne', 'ever', 'produced'), ('ever', 'produced', 'spanning'), ('produced', 'spanning', 'close'), ('spanning', 'clo
se', 'to'), ('close', 'to', 'four'), ('to', 'four', 'years'), ('four', 'years', 'and'), ('years', 'and', 'availa
ble'), ('and', 'available', 'in'), ('available', 'in', 'some'), ('in', 'some', 'developing'), ('some', 'developi
ng', 'countries'), ('developing', 'countries', 'until'), ('countries', 'until', 'early')], 'tesla cybertruck': [
('the', 'tesla', 'cybertruck'), ('tesla', 'cybertruck', 'is'), ('cybertruck', 'is', 'a'), ('is', 'a', 'pickup'),
('a', 'pickup', 'truck'), ('pickup', 'truck', 'manufactured'), ('truck', 'manufactured', 'by'), ('manufactured',
'by', 'tesla'), ('by', 'tesla', 'since'), ('tesla', 'since', 'it'), ('since', 'it', 'was'), ('it', 'was', 'first
'), ('was', 'first', 'unveiled'), ('first', 'unveiled', 'as'), ('unveiled', 'as', 'a'), ('as', 'a', 'prototype')
, ('a', 'prototype', 'in'), ('prototype', 'in', 'november'), ('in', 'november', 'featuring'), ('november', 'feat
uring', 'a'), ('featuring', 'a', 'distinctive'), ('a', 'distinctive', 'angular'), ('distinctive', 'angular', 'de
sign'), ('angular', 'design', 'composed'), ('design', 'composed', 'of'), ('composed', 'of', 'flat'), ('of', 'fla
t', 'unpainted'), ('flat', 'unpainted', 'stainless'), ('unpainted', 'stainless', 'steel'), ('stainless', 'steel'
, 'body'), ('steel', 'body', 'panels'), ('body', 'panels', 'drawing'), ('panels', 'drawing', 'comparisons'), ('d
rawing', 'comparisons', 'to'), ('comparisons', 'to', 'computer'), ('to', 'computer', 'models'), ('computer', 'mo
dels', 'originally'), ('models', 'originally', 'scheduled'), ('originally', 'scheduled', 'for'), ('scheduled', '
for', 'production'), ('for', 'production', 'in'), ('production', 'in', 'late'), ('in', 'late', 'the'), ('late',
'the', 'vehicle'), ('the', 'vehicle', 'faced'), ('vehicle', 'faced', 'multiple'), ('faced', 'multiple', 'delays'
), ('multiple', 'delays', 'before'), ('delays', 'before', 'entering'), ('before', 'entering', 'limited'), ('ente
ring', 'limited', 'production'), ('limited', 'production', 'at'), ('production', 'at', 'gigafactory'), ('at', 'g
igafactory', 'texas'), ('gigafactory', 'texas', 'in'), ('texas', 'in', 'november'), ('in', 'november', 'with'),
('november', 'with', 'initial'), ('with', 'initial', 'customer'), ('initial', 'customer', 'deliveries'), ('custo
mer', 'deliveries', 'occurring'), ('deliveries', 'occurring', 'later'), ('occurring', 'later', 'that'), ('later'
, 'that', 'month'), ('that', 'month', 'as'), ('month', 'as', 'of'), ('as', 'of', 'three'), ('of', 'three', 'vari
ants'), ('three', 'variants', 'are'), ('variants', 'are', 'available'), ('are', 'available', 'a'), ('available',
'a', 'drive'), ('a', 'drive', 'awd'), ('drive', 'awd', 'model'), ('awd', 'model', 'marketed'), ('model', 'market
ed', 'as'), ('marketed', 'as', 'the'), ('as', 'the', 'cyberbeast'), ('the', 'cyberbeast', 'a'), ('cyberbeast', '
a', 'awd'), ('a', 'awd', 'model'), ('awd', 'model', 'and'), ('model', 'and', 'a'), ('and', 'a', 'drive'), ('a',
'drive', 'rwd'), ('drive', 'rwd', 'long'), ('rwd', 'long', 'range'), ('long', 'range', 'model'), ('range', 'mode
l', 'epa'), ('model', 'epa', 'range'), ('epa', 'range', 'estimates'), ('range', 'estimates', 'vary'), ('estimate
s', 'vary', 'by'), ('vary', 'by', 'configuration'), ('by', 'configuration', 'from'), ('configuration', 'from', '
to'), ('from', 'to', 'miles'), ('to', 'miles', 'to'), ('miles', 'to', 'km'), ('to', 'km', 'as'), ('km', 'as', 'o
f'), ('as', 'of', 'the'), ('of', 'the', 'cybertruck'), ('the', 'cybertruck', 'is'), ('cybertruck', 'is', 'sold')
, ('is', 'sold', 'in'), ('sold', 'in', 'the'), ('in', 'the', 'united'), ('the', 'united', 'states'), ('united',
'states', 'mexico'), ('states', 'mexico', 'canada'), ('mexico', 'canada', 'and'), ('canada', 'and', 'south'), ('
and', 'south', 'korea'), ('south', 'korea', 'the'), ('korea', 'the', 'cybertruck'), ('the', 'cybertruck', 'has')
, ('cybertruck', 'has', 'been'), ('has', 'been', 'criticized'), ('been', 'criticized', 'for'), ('criticized', 'f
or', 'its'), ('for', 'its', 'production'), ('its', 'production', 'quality'), ('production', 'quality', 'and'), (
'quality', 'and', 'safety'), ('and', 'safety', 'concerns'), ('safety', 'concerns', 'while'), ('concerns', 'while
', 'its'), ('while', 'its', 'sales'), ('its', 'sales', 'have'), ('sales', 'have', 'been'), ('have', 'been', 'des
cribed'), ('been', 'described', 'as'), ('described', 'as', 'disappointing')], 'tesla model s': [('tesla', 'or',
'is'), ('or', 'is', 'an'), ('is', 'an', 'american'), ('an', 'american', 'multinational'), ('american', 'multinat
ional', 'automotive'), ('multinational', 'automotive', 'and'), ('automotive', 'and', 'clean'), ('and', 'clean',
'energy'), ('clean', 'energy', 'company'), ('energy', 'company', 'headquartered'), ('company', 'headquartered',
'in'), ('headquartered', 'in', 'austin'), ('in', 'austin', 'texas'), ('austin', 'texas', 'it'), ('texas', 'it',
'designs'), ('it', 'designs', 'manufactures'), ('designs', 'manufactures', 'and'), ('manufactures', 'and', 'sell
s'), ('and', 'sells', 'battery'), ('sells', 'battery', 'electric'), ('battery', 'electric', 'vehicles'), ('elect
ric', 'vehicles', 'bevs'), ('vehicles', 'bevs', 'stationary'), ('bevs', 'stationary', 'battery'), ('stationary',
'battery', 'energy'), ('battery', 'energy', 'storage'), ('energy', 'storage', 'devices'), ('storage', 'devices',
'from'), ('devices', 'from', 'home'), ('from', 'home', 'to'), ('home', 'to', 'solar'), ('to', 'solar', 'panels')
, ('solar', 'panels', 'and'), ('panels', 'and', 'solar'), ('and', 'solar', 'shingles'), ('solar', 'shingles', 'a
nd'), ('shingles', 'and', 'related'), ('and', 'related', 'products'), ('related', 'products', 'and'), ('products
', 'and', 'services'), ('and', 'services', 'tesla'), ('services', 'tesla', 'was'), ('tesla', 'was', 'incorporate
d'), ('was', 'incorporated', 'in'), ('incorporated', 'in', 'july'), ('in', 'july', 'by'), ('july', 'by', 'martin
'), ('by', 'martin', 'eberhard'), ('martin', 'eberhard', 'and'), ('eberhard', 'and', 'marc'), ('and', 'marc', 't
arpenning'), ('marc', 'tarpenning', 'as'), ('tarpenning', 'as', 'tesla'), ('as', 'tesla', 'motors'), ('tesla', '
motors', 'its'), ('motors', 'its', 'name'), ('its', 'name', 'is'), ('name', 'is', 'a'), ('is', 'a', 'tribute'),
('a', 'tribute', 'to'), ('tribute', 'to', 'the'), ('to', 'the', 'inventor'), ('the', 'inventor', 'and'), ('inven
tor', 'and', 'electrical'), ('and', 'electrical', 'engineer'), ('electrical', 'engineer', 'nikola'), ('engineer'
, 'nikola', 'tesla'), ('nikola', 'tesla', 'in'), ('tesla', 'in', 'february'), ('in', 'february', 'elon'), ('febr
uary', 'elon', 'musk'), ('elon', 'musk', 'led'), ('musk', 'led', 'tesla'), ('led', 'tesla', 'first'), ('tesla',
'first', 'funding'), ('first', 'funding', 'round'), ('funding', 'round', 'and'), ('round', 'and', 'became'), ('a
nd', 'became', 'the'), ('became', 'the', 'company'), ('the', 'company', 'chairman'), ('company', 'chairman', 'su
bsequently'), ('chairman', 'subsequently', 'claiming'), ('subsequently', 'claiming', 'to'), ('claiming', 'to', '
be'), ('to', 'be', 'a'), ('be', 'a', 'in'), ('a', 'in', 'he'), ('in', 'he', 'was'), ('he', 'was', 'named'), ('wa
s', 'named', 'chief'), ('named', 'chief', 'executive'), ('chief', 'executive', 'officer'), ('executive', 'office
r', 'in'), ('officer', 'in', 'the'), ('in', 'the', 'company'), ('the', 'company', 'began'), ('company', 'began',
'production'), ('began', 'production', 'of'), ('production', 'of', 'its'), ('of', 'its', 'first'), ('its', 'firs
t', 'car'), ('first', 'car', 'model'), ('car', 'model', 'the'), ('model', 'the', 'roadster'), ('the', 'roadster'
, 'sports'), ('roadster', 'sports', 'car'), ('sports', 'car', 'this'), ('car', 'this', 'was'), ('this', 'was', '
followed'), ('was', 'followed', 'by'), ('followed', 'by', 'the'), ('by', 'the', 'model'), ('the', 'model', 's'),
('model', 's', 'sedan'), ('s', 'sedan', 'in'), ('sedan', 'in', 'the'), ('in', 'the', 'model'), ('the', 'model',
'x'), ('model', 'x', 'suv'), ('x', 'suv', 'in'), ('suv', 'in', 'the'), ('in', 'the', 'model'), ('the', 'model',
'sedan'), ('model', 'sedan', 'in'), ('sedan', 'in', 'the'), ('in', 'the', 'model'), ('the', 'model', 'y'), ('mod
el', 'y', 'crossover'), ('y', 'crossover', 'in'), ('crossover', 'in', 'the'), ('in', 'the', 'tesla'), ('the', 't
esla', 'semi'), ('tesla', 'semi', 'truck'), ('semi', 'truck', 'in'), ('truck', 'in', 'and'), ('in', 'and', 'the'
), ('and', 'the', 'cybertruck'), ('the', 'cybertruck', 'pickup'), ('cybertruck', 'pickup', 'truck'), ('pickup',
'truck', 'in'), ('truck', 'in', 'tesla'), ('in', 'tesla', 'is'), ('tesla', 'is', 'one'), ('is', 'one', 'of'), ('
one', 'of', 'the'), ('of', 'the', 'world'), ('the', 'world', 'most'), ('world', 'most', 'valuable'), ('most', 'v
aluable', 'companies'), ('valuable', 'companies', 'in'), ('companies', 'in', 'terms'), ('in', 'terms', 'of'), ('
terms', 'of', 'market'), ('of', 'market', 'capitalization'), ('market', 'capitalization', 'starting'), ('capital
ization', 'starting', 'in'), ('starting', 'in', 'july'), ('in', 'july', 'it'), ('july', 'it', 'has'), ('it', 'ha
s', 'been'), ('has', 'been', 'the'), ('been', 'the', 'world'), ('the', 'world', 'most'), ('world', 'most', 'valu
able'), ('most', 'valuable', 'automaker'), ('valuable', 'automaker', 'from'), ('automaker', 'from', 'october'),
('from', 'october', 'to'), ('october', 'to', 'march'), ('to', 'march', 'tesla'), ('march', 'tesla', 'was'), ('te
sla', 'was', 'a'), ('was', 'a', 'company'), ('a', 'company', 'the'), ('company', 'the', 'seventh'), ('the', 'sev
enth', 'us'), ('seventh', 'us', 'company'), ('us', 'company', 'to'), ('company', 'to', 'reach'), ('to', 'reach',
'that'), ('reach', 'that', 'valuation'), ('that', 'valuation', 'tesla'), ('valuation', 'tesla', 'exceeded'), ('t
esla', 'exceeded', 'trillion'), ('exceeded', 'trillion', 'in'), ('trillion', 'in', 'market'), ('in', 'market', '
capitalization'), ('market', 'capitalization', 'again'), ('capitalization', 'again', 'between'), ('again', 'betw
een', 'november'), ('between', 'november', 'and'), ('november', 'and', 'february'), ('and', 'february', 'in'), (
'february', 'in', 'the'), ('in', 'the', 'company'), ('the', 'company', 'led'), ('company', 'led', 'the'), ('led'
, 'the', 'battery'), ('the', 'battery', 'electric'), ('battery', 'electric', 'vehicle'), ('electric', 'vehicle',
'market'), ('vehicle', 'market', 'with'), ('market', 'with', 'share'), ('with', 'share', 'in'), ('share', 'in',
'the'), ('in', 'the', 'company'), ('the', 'company', 'was'), ('company', 'was', 'ranked'), ('was', 'ranked', 'in
'), ('ranked', 'in', 'the'), ('in', 'the', 'forbes'), ('the', 'forbes', 'global'), ('forbes', 'global', 'in'), (
'global', 'in', 'may'), ('in', 'may', 'tesla'), ('may', 'tesla', 'again'), ('tesla', 'again', 'reached'), ('agai
n', 'reached', 'a'), ('reached', 'a', 'valuation'), ('a', 'valuation', 'of'), ('valuation', 'of', 'trillion'), (
'of', 'trillion', 'and'), ('trillion', 'and', 'has'), ('and', 'has', 'been'), ('has', 'been', 'a'), ('been', 'a'
, 'company'), ('a', 'company', 'since'), ('company', 'since', 'in'), ('since', 'in', 'november'), ('in', 'novemb
er', 'tesla'), ('november', 'tesla', 'approved'), ('tesla', 'approved', 'a'), ('approved', 'a', 'pay'), ('a', 'p
ay', 'package'), ('pay', 'package', 'worth'), ('package', 'worth', 'trillion'), ('worth', 'trillion', 'for'), ('
trillion', 'for', 'musk'), ('for', 'musk', 'which'), ('musk', 'which', 'he'), ('which', 'he', 'is'), ('he', 'is'
, 'to'), ('is', 'to', 'receive'), ('to', 'receive', 'over'), ('receive', 'over', 'years'), ('over', 'years', 'if
'), ('years', 'if', 'he'), ('if', 'he', 'meets'), ('he', 'meets', 'specific'), ('meets', 'specific', 'goals'), (
'specific', 'goals', 'tesla'), ('goals', 'tesla', 'has'), ('tesla', 'has', 'been'), ('has', 'been', 'the'), ('be
en', 'the', 'subject'), ('the', 'subject', 'of'), ('subject', 'of', 'lawsuits'), ('of', 'lawsuits', 'boycotts'),
('lawsuits', 'boycotts', 'government'), ('boycotts', 'government', 'scrutiny'), ('government', 'scrutiny', 'and'
), ('scrutiny', 'and', 'journalistic'), ('and', 'journalistic', 'criticism'), ('journalistic', 'criticism', 'ste
mming'), ('criticism', 'stemming', 'from'), ('stemming', 'from', 'allegations'), ('from', 'allegations', 'of'),
('allegations', 'of', 'multiple'), ('of', 'multiple', 'cases'), ('multiple', 'cases', 'of'), ('cases', 'of', 'wh
istleblower'), ('of', 'whistleblower', 'retaliation'), ('whistleblower', 'retaliation', 'worker'), ('retaliation
', 'worker', 'rights'), ('worker', 'rights', 'violations'), ('rights', 'violations', 'such'), ('violations', 'su
ch', 'as'), ('such', 'as', 'sexual'), ('as', 'sexual', 'harassment'), ('sexual', 'harassment', 'and'), ('harassm
ent', 'and', 'activities'), ('and', 'activities', 'safety'), ('activities', 'safety', 'defects'), ('safety', 'de
fects', 'leading'), ('defects', 'leading', 'to'), ('leading', 'to', 'dozens'), ('to', 'dozens', 'of'), ('dozens'
, 'of', 'recalls'), ('of', 'recalls', 'the'), ('recalls', 'the', 'lack'), ('the', 'lack', 'of'), ('lack', 'of',
'a'), ('of', 'a', 'public'), ('a', 'public', 'relations'), ('public', 'relations', 'department'), ('relations',
'department', 'and'), ('department', 'and', 'controversial'), ('and', 'controversial', 'statements'), ('controve
rsial', 'statements', 'from'), ('statements', 'from', 'musk'), ('from', 'musk', 'including'), ('musk', 'includin
g', 'overpromising'), ('including', 'overpromising', 'on'), ('overpromising', 'on', 'the'), ('on', 'the', 'compa
ny'), ('the', 'company', 'driving'), ('company', 'driving', 'assist'), ('driving', 'assist', 'technology'), ('as
sist', 'technology', 'and'), ('technology', 'and', 'product'), ('and', 'product', 'release'), ('product', 'relea
se', 'timelines')], 'playstation 5': [('the', 'playstation', 'is'), ('playstation', 'is', 'a'), ('is', 'a', 'hom
```

```
e'), ('a', 'home', 'video'), ('home', 'video', 'game'), ('video', 'game', 'console'), ('game', 'console', 'devel
oped'), ('console', 'developed', 'and'), ('developed', 'and', 'marketed'), ('and', 'marketed', 'by'), ('marketed
', 'by', 'sony'), ('by', 'sony', 'computer'), ('sony', 'computer', 'entertainment'), ('computer', 'entertainment
', 'it'), ('entertainment', 'it', 'was'), ('it', 'was', 'first'), ('was', 'first', 'released'), ('first', 'relea
sed', 'in'), ('released', 'in', 'japan'), ('in', 'japan', 'on'), ('japan', 'on', 'march'), ('on', 'march', 'in')
, ('march', 'in', 'north'), ('in', 'north', 'america'), ('north', 'america', 'on'), ('america', 'on', 'october')
, ('on', 'october', 'in'), ('october', 'in', 'europe'), ('in', 'europe', 'on'), ('europe', 'on', 'november'), ('
on', 'november', 'in'), ('november', 'in', 'australia'), ('in', 'australia', 'on'), ('australia', 'on', 'novembe
r'), ('on', 'november', 'and'), ('november', 'and', 'other'), ('and', 'other', 'regions'), ('other', 'regions', 
'thereafter'), ('regions', 'thereafter', 'it'), ('thereafter', 'it', 'is'), ('it', 'is', 'the'), ('is', 'the', '
successor'), ('the', 'successor', 'to'), ('successor', 'to', 'the'), ('to', 'the', 'original'), ('the', 'origina
l', 'playstation'), ('original', 'playstation', 'as'), ('playstation', 'as', 'well'), ('as', 'well', 'as'), ('we
ll', 'as', 'the'), ('as', 'the', 'second'), ('the', 'second', 'installment'), ('second', 'installment', 'in'), (
'installment', 'in', 'the'), ('in', 'the', 'playstation'), ('the', 'playstation', 'brand'), ('playstation', 'bra
nd', 'of'), ('brand', 'of', 'consoles'), ('of', 'consoles', 'as'), ('consoles', 'as', 'a'), ('as', 'a', 'console
'), ('a', 'console', 'it'), ('console', 'it', 'competed'), ('it', 'competed', 'with'), ('competed', 'with', 'nin
tendo'), ('with', 'nintendo', 'gamecube'), ('nintendo', 'gamecube', 'sega'), ('gamecube', 'sega', 'dreamcast'), 
('sega', 'dreamcast', 'and'), ('dreamcast', 'and', 'microsoft'), ('and', 'microsoft', 'xbox'), ('microsoft', 'xb
ox', 'announced'), ('xbox', 'announced', 'in'), ('announced', 'in', 'sony'), ('in', 'sony', 'began'), ('sony', '
began', 'developing'), ('began', 'developing', 'the'), ('developing', 'the', 'console'), ('the', 'console', 'aft
er'), ('console', 'after', 'the'), ('after', 'the', 'immense'), ('the', 'immense', 'success'), ('immense', 'succ
ess', 'of'), ('success', 'of', 'its'), ('of', 'its', 'predecessor'), ('its', 'predecessor', 'in'), ('predecessor
', 'in', 'addition'), ('in', 'addition', 'to'), ('addition', 'to', 'serving'), ('to', 'serving', 'as'), ('servin
g', 'as', 'a'), ('as', 'a', 'game'), ('a', 'game', 'console'), ('game', 'console', 'it'), ('console', 'it', 'fea
tures'), ('it', 'features', 'a'), ('features', 'a', 'dvd'), ('a', 'dvd', 'drive'), ('dvd', 'drive', 'and'), ('dr
ive', 'and', 'was'), ('and', 'was', 'priced'), ('was', 'priced', 'lower'), ('priced', 'lower', 'than'), ('lower'
, 'than', 'standalone'), ('than', 'standalone', 'dvd'), ('standalone', 'dvd', 'players'), ('dvd', 'players', 'of
'), ('players', 'of', 'the'), ('of', 'the', 'time'), ('the', 'time', 'enhancing'), ('time', 'enhancing', 'its'),
('enhancing', 'its', 'value'), ('its', 'value', 'full'), ('value', 'full', 'backward'), ('full', 'backward', 'co
mpatibility'), ('backward', 'compatibility', 'with'), ('compatibility', 'with', 'original'), ('with', 'original'
, 'playstation'), ('original', 'playstation', 'games'), ('playstation', 'games', 'and'), ('games', 'and', 'acces
sories'), ('and', 'accessories', 'gave'), ('accessories', 'gave', 'it'), ('gave', 'it', 'access'), ('it', 'acces
s', 'to'), ('access', 'to', 'a'), ('to', 'a', 'vast'), ('a', 'vast', 'launch'), ('vast', 'launch', 'library'), (
'launch', 'library', 'far'), ('library', 'far', 'surpassing'), ('far', 'surpassing', 'those'), ('surpassing', 't
hose', 'of'), ('those', 'of', 'its'), ('of', 'its', 'competitors'), ('its', 'competitors', 'the'), ('competitors
', 'the', 'console'), ('the', 'console', 'hardware'), ('console', 'hardware', 'was'), ('hardware', 'was', 'also'
), ('was', 'also', 'notable'), ('also', 'notable', 'for'), ('notable', 'for', 'its'), ('for', 'its', 'emotion'),
('its', 'emotion', 'engine'), ('emotion', 'engine', 'processor'), ('engine', 'processor', 'with'), ('processor',
'with', 'toshiba'), ('with', 'toshiba', 'which'), ('toshiba', 'which', 'was'), ('which', 'was', 'promoted'), ('w
as', 'promoted', 'as'), ('promoted', 'as', 'being'), ('as', 'being', 'more'), ('being', 'more', 'powerful'), ('m
ore', 'powerful', 'than'), ('powerful', 'than', 'most'), ('than', 'most', 'personal'), ('most', 'personal', 'com
puters'), ('personal', 'computers', 'of'), ('computers', 'of', 'the'), ('of', 'the', 'era'), ('the', 'era', 'the
'), ('era', 'the', 'playstation'), ('the', 'playstation', 'remains'), ('playstation', 'remains', 'the'), ('remai
ns', 'the', 'video'), ('the', 'video', 'game'), ('video', 'game', 'console'), ('game', 'console', 'of'), ('conso
le', 'of', 'all'), ('of', 'all', 'time'), ('all', 'time', 'having'), ('time', 'having', 'sold'), ('having', 'sol
d', 'million'), ('sold', 'million', 'units'), ('million', 'units', 'worldwide'), ('units', 'worldwide', 'nearly'
), ('worldwide', 'nearly', 'triple'), ('nearly', 'triple', 'the'), ('triple', 'the', 'combined'), ('the', 'combi
ned', 'sales'), ('combined', 'sales', 'of'), ('sales', 'of', 'competing'), ('of', 'competing', 'consoles'), ('co
mpeting', 'consoles', 'it'), ('consoles', 'it', 'received'), ('it', 'received', 'widespread'), ('received', 'wid
espread', 'critical'), ('widespread', 'critical', 'acclaim'), ('critical', 'acclaim', 'and'), ('acclaim', 'and',
'amassed'), ('and', 'amassed', 'a'), ('amassed', 'a', 'global'), ('a', 'global', 'library'), ('global', 'library
', 'of'), ('library', 'of', 'game'), ('of', 'game', 'titles'), ('game', 'titles', 'with'), ('titles', 'with', 'b
illion'), ('with', 'billion', 'copies'), ('billion', 'copies', 'sold'), ('copies', 'sold', 'in'), ('sold', 'in',
'sony'), ('in', 'sony', 'revised'), ('sony', 'revised', 'the'), ('revised', 'the', 'console'), ('the', 'console'
, 'with'), ('console', 'with', 'a'), ('with', 'a', 'smaller'), ('a', 'smaller', 'lighter'), ('smaller', 'lighter
', 'body'), ('lighter', 'body', 'officially'), ('body', 'officially', 'known'), ('officially', 'known', 'as'), (
'known', 'as', 'the'), ('as', 'the', 'slimline'), ('the', 'slimline', 'even'), ('slimline', 'even', 'after'), ('
even', 'after', 'the'), ('after', 'the', 'release'), ('the', 'release', 'of'), ('release', 'of', 'its'), ('of', 
'its', 'successor'), ('its', 'successor', 'the'), ('successor', 'the', 'playstation'), ('the', 'playstation', 'i
n'), ('playstation', 'in', 'it'), ('in', 'it', 'remained'), ('it', 'remained', 'in'), ('remained', 'in', 'produc
tion'), ('in', 'production', 'and'), ('production', 'and', 'continued'), ('and', 'continued', 'to'), ('continued
', 'to', 'receive'), ('to', 'receive', 'new'), ('receive', 'new', 'game'), ('new', 'game', 'releases'), ('game',
'releases', 'for'), ('releases', 'for', 'several'), ('for', 'several', 'years'), ('several', 'years', 'with'), (
'years', 'with', 'the'), ('with', 'the', 'last'), ('the', 'last', 'game'), ('last', 'game', 'for'), ('game', 'fo
r', 'the'), ('for', 'the', 'system'), ('the', 'system', 'pro'), ('system', 'pro', 'evolution'), ('pro', 'evoluti
on', 'soccer'), ('evolution', 'soccer', 'being'), ('soccer', 'being', 'released'), ('being', 'released', 'in'), 
('released', 'in', 'europe'), ('in', 'europe', 'in'), ('europe', 'in', 'november'), ('in', 'november', 'manufact
uring'), ('november', 'manufacturing', 'officially'), ('manufacturing', 'officially', 'ended'), ('officially', '
ended', 'in'), ('ended', 'in', 'early'), ('in', 'early', 'giving'), ('early', 'giving', 'the'), ('giving', 'the'
, 'console'), ('the', 'console', 'one'), ('console', 'one', 'of'), ('one', 'of', 'the'), ('of', 'the', 'longest'
), ('the', 'longest', 'lifespans'), ('longest', 'lifespans', 'in'), ('lifespans', 'in', 'video'), ('in', 'video'
, 'game'), ('video', 'game', 'history')], 'iphone 6': [('the', 'iphone', 'is'), ('iphone', 'is', 'a'), ('is', 'a
', 'smartphone'), ('a', 'smartphone', 'that'), ('smartphone', 'that', 'was'), ('that', 'was', 'developed'), ('wa
s', 'developed', 'and'), ('developed', 'and', 'marketed'), ('and', 'marketed', 'by'), ('marketed', 'by', 'apple'
), ('by', 'apple', 'it'), ('apple', 'it', 'is'), ('it', 'is', 'the'), ('is', 'the', 'generation'), ('the', 'gene
ration', 'iphone'), ('generation', 'iphone', 'succeeding'), ('iphone', 'succeeding', 'the'), ('succeeding', 'the
', 'iphone'), ('the', 'iphone', 'and'), ('iphone', 'and', 'preceding'), ('and', 'preceding', 'both'), ('precedin
g', 'both', 'the'), ('both', 'the', 'iphone'), ('the', 'iphone', 'and'), ('iphone', 'and', 'iphone'), ('and', 'i
phone', 'it'), ('iphone', 'it', 'was'), ('it', 'was', 'formally'), ('was', 'formally', 'unveiled'), ('formally',
'unveiled', 'as'), ('unveiled', 'as', 'part'), ('as', 'part', 'of'), ('part', 'of', 'a'), ('of', 'a', 'press'), 
('a', 'press', 'event'), ('press', 'event', 'on'), ('event', 'on', 'september'), ('on', 'september', 'and'), ('s
eptember', 'and', 'subsequently'), ('and', 'subsequently', 'released'), ('subsequently', 'released', 'on'), ('re
leased', 'on', 'september'), ('on', 'september', 'the'), ('september', 'the', 'iphone'), ('the', 'iphone', 'was'
```

```
), ('iphone', 'was', 'the'), ('was', 'the', 'first'), ('the', 'first', 'iphone'), ('first', 'iphone', 'to'), ('i
phone', 'to', 'be'), ('to', 'be', 'announced'), ('be', 'announced', 'in'), ('announced', 'in', 'september'), ('i
n', 'september', 'and'), ('september', 'and', 'setting'), ('and', 'setting', 'a'), ('setting', 'a', 'trend'), ('
a', 'trend', 'for'), ('trend', 'for', 'subsequent'), ('for', 'subsequent', 'iphone'), ('subsequent', 'iphone', '
releases'), ('iphone', 'releases', 'the'), ('releases', 'the', 'first'), ('the', 'first', 'iphone'), ('first', '
iphone', 'to'), ('iphone', 'to', 'be'), ('to', 'be', 'completely'), ('be', 'completely', 'developed'), ('complet
ely', 'developed', 'under'), ('developed', 'under', 'the'), ('under', 'the', 'guidance'), ('the', 'guidance', 'o
f'), ('guidance', 'of', 'tim'), ('of', 'tim', 'cook'), ('tim', 'cook', 'and'), ('cook', 'and', 'the'), ('and', '
the', 'last'), ('the', 'last', 'iphone'), ('last', 'iphone', 'to'), ('iphone', 'to', 'be'), ('to', 'be', 'overse
en'), ('be', 'overseen', 'by'), ('overseen', 'by', 'steve'), ('by', 'steve', 'jobs'), ('steve', 'jobs', 'the'),
('jobs', 'the', 'iphone'), ('the', 'iphone', 'design'), ('iphone', 'design', 'was'), ('design', 'was', 'used'),
('was', 'used', 'three'), ('used', 'three', 'times'), ('three', 'times', 'first'), ('times', 'first', 'with'), (
'first', 'with', 'the'), ('with', 'the', 'iphone'), ('the', 'iphone', 'itself'), ('iphone', 'itself', 'in'), ('i
tself', 'in', 'then'), ('in', 'then', 'with'), ('then', 'with', 'the'), ('with', 'the', 'iphone'), ('the', 'ipho
ne', 'in'), ('iphone', 'in', 'and'), ('in', 'and', 'finally'), ('and', 'finally', 'with'), ('finally', 'with', '
the'), ('with', 'the', 'iphone'), ('the', 'iphone', 'se'), ('iphone', 'se', 'in'), ('se', 'in', 'the'), ('in', '
the', 'iphone'), ('the', 'iphone', 'featured'), ('iphone', 'featured', 'major'), ('featured', 'major', 'design')
, ('major', 'design', 'changes'), ('design', 'changes', 'in'), ('changes', 'in', 'comparison'), ('in', 'comparis
on', 'to'), ('comparison', 'to', 'its'), ('to', 'its', 'predecessor'), ('its', 'predecessor', 'these'), ('predec
essor', 'these', 'included'), ('these', 'included', 'an'), ('included', 'an', 'body'), ('an', 'body', 'which'),
('body', 'which', 'was'), ('which', 'was', 'thinner'), ('was', 'thinner', 'and'), ('thinner', 'and', 'lighter'),
('and', 'lighter', 'than'), ('lighter', 'than', 'previous'), ('than', 'previous', 'models'), ('previous', 'model
s', 'a'), ('models', 'a', 'taller'), ('a', 'taller', 'screen'), ('taller', 'screen', 'with'), ('screen', 'with',
'a'), ('with', 'a', 'nearly'), ('a', 'nearly', 'aspect'), ('nearly', 'aspect', 'ratio'), ('aspect', 'ratio', 'th
e'), ('ratio', 'the', 'apple'), ('the', 'apple', 'lte'), ('apple', 'lte', 'support'), ('lte', 'support', 'and'),
('support', 'and', 'lightning'), ('and', 'lightning', 'a'), ('lightning', 'a', 'new'), ('a', 'new', 'compact'),
('new', 'compact', 'dock'), ('compact', 'dock', 'connector'), ('dock', 'connector', 'which'), ('connector', 'whi
ch', 'replaced'), ('which', 'replaced', 'the'), ('replaced', 'the', 'design'), ('the', 'design', 'used'), ('desi
gn', 'used', 'by'), ('used', 'by', 'previous'), ('by', 'previous', 'iphone'), ('previous', 'iphone', 'models'),
('iphone', 'models', 'this'), ('models', 'this', 'was'), ('this', 'was', 'the'), ('was', 'the', 'second'), ('the
', 'second', 'iphone'), ('second', 'iphone', 'after'), ('iphone', 'after', 'the'), ('after', 'the', 'iphone'), (
'the', 'iphone', 'to'), ('iphone', 'to', 'include'), ('to', 'include', 'apple'), ('include', 'apple', 'new'), ('
apple', 'new', 'mp'), ('new', 'mp', 'camera'), ('mp', 'camera', 'apple'), ('camera', 'apple', 'began'), ('apple'
, 'began', 'taking'), ('began', 'taking', 'on'), ('taking', 'on', 'september'), ('on', 'september', 'and'), ('se
ptember', 'and', 'over'), ('and', 'over', 'two'), ('over', 'two', 'million'), ('two', 'million', 'were'), ('mill
ion', 'were', 'received'), ('were', 'received', 'within'), ('received', 'within', 'hours'), ('within', 'hours',
'initial'), ('hours', 'initial', 'demand'), ('initial', 'demand', 'for'), ('demand', 'for', 'the'), ('for', 'the
', 'iphone'), ('the', 'iphone', 'exceeded'), ('iphone', 'exceeded', 'the'), ('exceeded', 'the', 'supply'), ('the
', 'supply', 'available'), ('supply', 'available', 'at'), ('available', 'at', 'launch'), ('at', 'launch', 'on'),
('launch', 'on', 'september'), ('on', 'september', 'and'), ('september', 'and', 'was'), ('and', 'was', 'describe
d'), ('was', 'described', 'by'), ('described', 'by', 'apple'), ('by', 'apple', 'as'), ('apple', 'as', 'extraordi
nary'), ('as', 'extraordinary', 'with'), ('extraordinary', 'with', 'having'), ('with', 'having', 'sold'), ('havi
ng', 'sold', 'twenty'), ('sold', 'twenty', 'times'), ('twenty', 'times', 'faster'), ('times', 'faster', 'than'),
('faster', 'than', 'its'), ('than', 'its', 'predecessors'), ('its', 'predecessors', 'while'), ('predecessors', '
while', 'reception'), ('while', 'reception', 'to'), ('reception', 'to', 'the'), ('to', 'the', 'iphone'), ('the',
'iphone', 'was'), ('iphone', 'was', 'generally'), ('was', 'generally', 'positive'), ('generally', 'positive', 'c
onsumers'), ('positive', 'consumers', 'and'), ('consumers', 'and', 'reviewers'), ('and', 'reviewers', 'noted'),
('reviewers', 'noted', 'hardware'), ('noted', 'hardware', 'issues'), ('hardware', 'issues', 'such'), ('issues',
'such', 'as'), ('such', 'as', 'an'), ('as', 'an', 'unintended'), ('an', 'unintended', 'purple'), ('unintended',
'purple', 'hue'), ('purple', 'hue', 'in'), ('hue', 'in', 'photos'), ('in', 'photos', 'taken'), ('photos', 'taken
', 'and'), ('taken', 'and', 'the'), ('and', 'the', 'phone'), ('the', 'phone', 'coating'), ('phone', 'coating', '
being'), ('coating', 'being', 'prone'), ('being', 'prone', 'to'), ('prone', 'to', 'chipping'), ('to', 'chipping'
, 'reception'), ('chipping', 'reception', 'was'), ('reception', 'was', 'also'), ('was', 'also', 'mixed'), ('also
', 'mixed', 'over'), ('mixed', 'over', 'apple'), ('over', 'apple', 'decision'), ('apple', 'decision', 'to'), ('d
ecision', 'to', 'switch'), ('to', 'switch', 'to'), ('switch', 'to', 'a'), ('to', 'a', 'different'), ('a', 'diffe
rent', 'dock'), ('different', 'dock', 'connector'), ('dock', 'connector', 'design'), ('connector', 'design', 'as
'), ('design', 'as', 'the'), ('as', 'the', 'change'), ('the', 'change', 'affected'), ('change', 'affected', 'iph
one'), ('affected', 'iphone', 'compatibility'), ('iphone', 'compatibility', 'with'), ('compatibility', 'with', '
accessories'), ('with', 'accessories', 'that'), ('accessories', 'that', 'were'), ('that', 'were', 'otherwise'),
('were', 'otherwise', 'compatible'), ('otherwise', 'compatible', 'with'), ('compatible', 'with', 'previous'), ('
with', 'previous', 'iterations'), ('previous', 'iterations', 'of'), ('iterations', 'of', 'the'), ('of', 'the', '
line'), ('the', 'line', 'alongside'), ('line', 'alongside', 'the'), ('alongside', 'the', 'iphone'), ('the', 'iph
one', 'the'), ('iphone', 'the', 'iphone'), ('the', 'iphone', 'was'), ('iphone', 'was', 'officially'), ('was', 'o
fficially', 'discontinued'), ('officially', 'discontinued', 'by'), ('discontinued', 'by', 'apple'), ('by', 'appl
e', 'on'), ('apple', 'on', 'september'), ('on', 'september', 'with'), ('september', 'with', 'the'), ('with', 'th
e', 'announcement'), ('the', 'announcement', 'of'), ('announcement', 'of', 'its'), ('of', 'its', 'successors'),
('its', 'successors', 'the'), ('successors', 'the', 'iphone'), ('the', 'iphone', 'and'), ('iphone', 'and', 'the'
), ('and', 'the', 'iphone'), ('the', 'iphone', 'the'), ('iphone', 'the', 'iphone'), ('the', 'iphone', 'has'), ('
iphone', 'has', 'the'), ('has', 'the', 'joint'), ('the', 'joint', 'lifespan'), ('joint', 'lifespan', 'of'), ('li
fespan', 'of', 'any'), ('of', 'any', 'iphone'), ('any', 'iphone', 'ever'), ('iphone', 'ever', 'produced'), ('eve
r', 'produced', 'with'), ('produced', 'with', 'only'), ('with', 'only', 'twelve'), ('only', 'twelve', 'months'),
('twelve', 'months', 'in'), ('months', 'in', 'production'), ('in', 'production', 'breaking'), ('production', 'br
eaking', 'with'), ('breaking', 'with', 'apple'), ('with', 'apple', 'standard'), ('apple', 'standard', 'practice'
), ('standard', 'practice', 'of'), ('practice', 'of', 'selling'), ('of', 'selling', 'an'), ('selling', 'an', 'ex
isting'), ('an', 'existing', 'iphone'), ('existing', 'iphone', 'model'), ('iphone', 'model', 'at'), ('model', 'a
t', 'a'), ('at', 'a', 'reduced'), ('a', 'reduced', 'price'), ('reduced', 'price', 'upon'), ('price', 'upon', 'th
e'), ('upon', 'the', 'release'), ('the', 'release', 'of'), ('release', 'of', 'a'), ('of', 'a', 'new'), ('a', 'ne
w', 'model'), ('new', 'model', 'this'), ('model', 'this', 'was'), ('this', 'was', 'broken'), ('was', 'broken', '
by'), ('broken', 'by', 'the'), ('by', 'the', 'iphone'), ('the', 'iphone', 'x'), ('iphone', 'x', 'which'), ('x',
'which', 'only'), ('which', 'only', 'had'), ('only', 'had', 'in'), ('had', 'in', 'production'), ('in', 'producti
on', 'from'), ('production', 'from', 'november'), ('from', 'november', 'to'), ('november', 'to', 'september'), (
'to', 'september', 'and'), ('september', 'and', 'tied'), ('and', 'tied', 'with'), ('tied', 'with', 'the'), ('wit
h', 'the', 'iphone'), ('the', 'iphone', 'xs'), ('iphone', 'xs', 'which'), ('xs', 'which', 'had'), ('which', 'had
```

```
', 'from'), ('had', 'from', 'september'), ('from', 'september', 'to'), ('september', 'to', 'september'), ('to',
'september', 'the'), ('september', 'the', 'iphone'), ('the', 'iphone', 'pro'), ('iphone', 'pro', 'and'), ('pro',
'and', 'subsequent'), ('and', 'subsequent', 'pro'), ('subsequent', 'pro', 'designated'), ('pro', 'designated', '
iphones'), ('designated', 'iphones', 'have'), ('iphones', 'have', 'also'), ('have', 'also', 'had'), ('also', 'ha
d', 'twelve'), ('had', 'twelve', 'month'), ('twelve', 'month', 'availability'), ('month', 'availability', 'being
'), ('availability', 'being', 'discontinued'), ('being', 'discontinued', 'upon'), ('discontinued', 'upon', 'rele
ase'), ('upon', 'release', 'of'), ('release', 'of', 'its'), ('of', 'its', 'successor'), ('its', 'successor', 'th
e'), ('successor', 'the', 'iphone'), ('the', 'iphone', 'was'), ('iphone', 'was', 'replaced'), ('was', 'replaced'
, 'as'), ('replaced', 'as', 'a'), ('as', 'a', 'midrange'), ('a', 'midrange', 'and'), ('midrange', 'and', 'then')
, ('and', 'then', 'an'), ('then', 'an', 'device'), ('an', 'device', 'by'), ('device', 'by', 'the'), ('by', 'the'
, 'iphone'), ('the', 'iphone', 'the'), ('iphone', 'the', 'internal'), ('the', 'internal', 'hardware'), ('interna
l', 'hardware', 'specifications'), ('hardware', 'specifications', 'are'), ('specifications', 'are', 'almost'), (
'are', 'almost', 'identical'), ('almost', 'identical', 'to'), ('identical', 'to', 'the'), ('to', 'the', 'albeit'
), ('the', 'albeit', 'having'), ('albeit', 'having', 'a'), ('having', 'a', 'less'), ('a', 'less', 'expensive'),
('less', 'expensive', 'polycarbonate'), ('expensive', 'polycarbonate', 'exterior'), ('polycarbonate', 'exterior'
, 'shell'), ('exterior', 'shell', 'the'), ('shell', 'the', 'iphone'), ('the', 'iphone', 'supports'), ('iphone',
'supports', 'ios'), ('supports', 'ios', 'and'), ('ios', 'and', 'the'), ('and', 'the', 'iphone'), ('the', 'iphone
', 'does'), ('iphone', 'does', 'not'), ('does', 'not', 'support'), ('not', 'support', 'ios'), ('support', 'ios',
'due'), ('ios', 'due', 'to'), ('due', 'to', 'it'), ('to', 'it', 'dropping'), ('it', 'dropping', 'support'), ('dr
opping', 'support', 'for'), ('support', 'for', 'devices'), ('for', 'devices', 'the'), ('devices', 'the', 'iphone
'), ('the', 'iphone', 'is'), ('iphone', 'is', 'the'), ('is', 'the', 'second'), ('the', 'second', 'iphone'), ('se
cond', 'iphone', 'to'), ('iphone', 'to', 'support'), ('to', 'support', 'five'), ('support', 'five', 'major'), ('
five', 'major', 'versions'), ('major', 'versions', 'of'), ('versions', 'of', 'ios'), ('of', 'ios', 'after'), ('i
os', 'after', 'the'), ('after', 'the', 'iphone')], 'playstation 4': [('the', 'playstation', 'codenamed'), ('play
station', 'codenamed', 'psx'), ('codenamed', 'psx', 'abbreviated'), ('psx', 'abbreviated', 'as'), ('abbreviated'
, 'as', 'ps'), ('as', 'ps', 'and'), ('ps', 'and', 'retroactively'), ('and', 'retroactively', 'or'), ('retroactiv
ely', 'or', 'ps'), ('or', 'ps', 'one'), ('ps', 'one', 'is'), ('one', 'is', 'a'), ('is', 'a', 'home'), ('a', 'hom
e', 'video'), ('home', 'video', 'game'), ('video', 'game', 'console'), ('game', 'console', 'developed'), ('conso
le', 'developed', 'and'), ('developed', 'and', 'marketed'), ('and', 'marketed', 'by'), ('marketed', 'by', 'sony'
), ('by', 'sony', 'computer'), ('sony', 'computer', 'entertainment'), ('computer', 'entertainment', 'it'), ('ent
ertainment', 'it', 'was'), ('it', 'was', 'released'), ('was', 'released', 'in'), ('released', 'in', 'japan'), ('
in', 'japan', 'on'), ('japan', 'on', 'december'), ('on', 'december', 'followed'), ('december', 'followed', 'by')
, ('followed', 'by', 'north'), ('by', 'north', 'america'), ('north', 'america', 'on'), ('america', 'on', 'septem
ber'), ('on', 'september', 'europe'), ('september', 'europe', 'on'), ('europe', 'on', 'september'), ('on', 'sept
ember', 'and'), ('september', 'and', 'other'), ('and', 'other', 'regions'), ('other', 'regions', 'following'), (
'regions', 'following', 'thereafter'), ('following', 'thereafter', 'as'), ('thereafter', 'as', 'a'), ('as', 'a',
'console'), ('a', 'console', 'the'), ('console', 'the', 'playstation'), ('the', 'playstation', 'primarily'), ('p
laystation', 'primarily', 'competed'), ('primarily', 'competed', 'with'), ('competed', 'with', 'the'), ('with',
'the', 'nintendo'), ('the', 'nintendo', 'and'), ('nintendo', 'and', 'the'), ('and', 'the', 'sega'), ('the', 'seg
a', 'saturn'), ('sega', 'saturn', 'sony'), ('saturn', 'sony', 'began'), ('sony', 'began', 'developing'), ('began
', 'developing', 'the'), ('developing', 'the', 'playstation'), ('the', 'playstation', 'after'), ('playstation',
'after', 'a'), ('after', 'a', 'failed'), ('a', 'failed', 'venture'), ('failed', 'venture', 'with'), ('venture',
'with', 'nintendo'), ('with', 'nintendo', 'to'), ('nintendo', 'to', 'create'), ('to', 'create', 'a'), ('create',
'a', 'peripheral'), ('a', 'peripheral', 'for'), ('peripheral', 'for', 'the'), ('for', 'the', 'super'), ('the', '
super', 'nintendo'), ('super', 'nintendo', 'entertainment'), ('nintendo', 'entertainment', 'system'), ('entertai
nment', 'system', 'in'), ('system', 'in', 'the'), ('in', 'the', 'early'), ('the', 'early', 'the'), ('early', 'th
e', 'console'), ('the', 'console', 'was'), ('console', 'was', 'primarily'), ('was', 'primarily', 'designed'), ('
primarily', 'designed', 'by'), ('designed', 'by', 'ken'), ('by', 'ken', 'kutaragi'), ('ken', 'kutaragi', 'and'),
('kutaragi', 'and', 'sony'), ('and', 'sony', 'computer'), ('sony', 'computer', 'entertainment'), ('computer', 'e
ntertainment', 'in'), ('entertainment', 'in', 'japan'), ('in', 'japan', 'while'), ('japan', 'while', 'additional
'), ('while', 'additional', 'development'), ('additional', 'development', 'was'), ('development', 'was', 'outsou
rced'), ('was', 'outsourced', 'in'), ('outsourced', 'in', 'the'), ('in', 'the', 'united'), ('the', 'united', 'ki
ngdom'), ('united', 'kingdom', 'an'), ('kingdom', 'an', 'emphasis'), ('an', 'emphasis', 'on'), ('emphasis', 'on'
, 'polygon'), ('on', 'polygon', 'graphics'), ('polygon', 'graphics', 'was'), ('graphics', 'was', 'placed'), ('wa
s', 'placed', 'at'), ('placed', 'at', 'the'), ('at', 'the', 'forefront'), ('the', 'forefront', 'of'), ('forefron
t', 'of', 'the'), ('of', 'the', 'console'), ('the', 'console', 'design'), ('console', 'design', 'playstation'),
('design', 'playstation', 'game'), ('playstation', 'game', 'production'), ('game', 'production', 'was'), ('produ
ction', 'was', 'designed'), ('was', 'designed', 'to'), ('designed', 'to', 'be'), ('to', 'be', 'streamlined'), ('
be', 'streamlined', 'and'), ('streamlined', 'and', 'inclusive'), ('and', 'inclusive', 'enticing'), ('inclusive',
'enticing', 'the'), ('enticing', 'the', 'support'), ('the', 'support', 'of'), ('support', 'of', 'many'), ('of',
'many', 'third'), ('many', 'third', 'party'), ('third', 'party', 'developers'), ('party', 'developers', 'the'),
('developers', 'the', 'console'), ('the', 'console', 'proved'), ('console', 'proved', 'popular'), ('proved', 'po
pular', 'for'), ('popular', 'for', 'its'), ('for', 'its', 'extensive'), ('its', 'extensive', 'game'), ('extensiv
e', 'game', 'library'), ('game', 'library', 'popular'), ('library', 'popular', 'franchises'), ('popular', 'franc
hises', 'low'), ('franchises', 'low', 'retail'), ('low', 'retail', 'price'), ('retail', 'price', 'and'), ('price
', 'and', 'aggressive'), ('and', 'aggressive', 'youth'), ('aggressive', 'youth', 'marketing'), ('youth', 'market
ing', 'which'), ('marketing', 'which', 'advertised'), ('which', 'advertised', 'it'), ('advertised', 'it', 'as'),
('it', 'as', 'the'), ('as', 'the', 'preferable'), ('the', 'preferable', 'console'), ('preferable', 'console', 'f
or'), ('console', 'for', 'adolescents'), ('for', 'adolescents', 'and'), ('adolescents', 'and', 'adults'), ('and'
, 'adults', 'critically'), ('adults', 'critically', 'acclaimed'), ('critically', 'acclaimed', 'games'), ('acclai
med', 'games', 'that'), ('games', 'that', 'defined'), ('that', 'defined', 'the'), ('defined', 'the', 'console'),
('the', 'console', 'include'), ('console', 'include', 'gran'), ('include', 'gran', 'turismo'), ('gran', 'turismo
', 'crash'), ('turismo', 'crash', 'bandicoot'), ('crash', 'bandicoot', 'spyro'), ('bandicoot', 'spyro', 'the'),
('spyro', 'the', 'dragon'), ('the', 'dragon', 'tomb'), ('dragon', 'tomb', 'raider'), ('tomb', 'raider', 'residen
t'), ('raider', 'resident', 'evil'), ('resident', 'evil', 'metal'), ('evil', 'metal', 'gear'), ('metal', 'gear',
'solid'), ('gear', 'solid', 'tekken'), ('solid', 'tekken', 'and'), ('tekken', 'and', 'final'), ('and', 'final',
'fantasy'), ('final', 'fantasy', 'vii'), ('fantasy', 'vii', 'sony'), ('vii', 'sony', 'ceased'), ('sony', 'ceased
', 'production'), ('ceased', 'production', 'of'), ('production', 'of', 'the'), ('of', 'the', 'playstation'), ('t
he', 'playstation', 'on'), ('playstation', 'on', 'march'), ('on', 'march', 'eleven'), ('march', 'eleven', 'years
'), ('eleven', 'years', 'after'), ('years', 'after', 'it'), ('after', 'it', 'had'), ('it', 'had', 'been'), ('had
', 'been', 'released'), ('been', 'released', 'and'), ('released', 'and', 'in'), ('and', 'in', 'the'), ('in', 'th
e', 'same'), ('the', 'same', 'year'), ('same', 'year', 'the'), ('year', 'the', 'playstation'), ('the', 'playstat
ion', 'debuted'), ('playstation', 'debuted', 'more'), ('debuted', 'more', 'than'), ('more', 'than', 'playstation
```

```
'), ('than', 'playstation', 'games'), ('playstation', 'games', 'were'), ('games', 'were', 'released'), ('were',
'released', 'with'), ('released', 'with', 'cumulative'), ('with', 'cumulative', 'sales'), ('cumulative', 'sales'
, 'of'), ('sales', 'of', 'million'), ('of', 'million', 'units'), ('million', 'units', 'the'), ('units', 'the', '
playstation'), ('the', 'playstation', 'signaled'), ('playstation', 'signaled', 'sony'), ('signaled', 'sony', 'ri
se'), ('sony', 'rise', 'to'), ('rise', 'to', 'power'), ('to', 'power', 'in'), ('power', 'in', 'the'), ('in', 'th
e', 'video'), ('the', 'video', 'game'), ('video', 'game', 'industry'), ('game', 'industry', 'it'), ('industry',
'it', 'received'), ('it', 'received', 'acclaim'), ('received', 'acclaim', 'and'), ('acclaim', 'and', 'sold'), ('
and', 'sold', 'strongly'), ('sold', 'strongly', 'in'), ('strongly', 'in', 'less'), ('in', 'less', 'than'), ('les
s', 'than', 'a'), ('than', 'a', 'decade'), ('a', 'decade', 'it'), ('decade', 'it', 'became'), ('it', 'became', '
the'), ('became', 'the', 'first'), ('the', 'first', 'computer'), ('first', 'computer', 'entertainment'), ('compu
ter', 'entertainment', 'platform'), ('entertainment', 'platform', 'to'), ('platform', 'to', 'ship'), ('to', 'shi
p', 'over'), ('ship', 'over', 'million'), ('over', 'million', 'units'), ('million', 'units', 'its'), ('units', '
its', 'use'), ('its', 'use', 'of'), ('use', 'of', 'compact'), ('of', 'compact', 'discs'), ('compact', 'discs', '
heralded'), ('discs', 'heralded', 'the'), ('heralded', 'the', 'game'), ('the', 'game', 'industry'), ('game', 'in
dustry', 'transition'), ('industry', 'transition', 'from'), ('transition', 'from', 'cartridges'), ('from', 'cart
ridges', 'the'), ('cartridges', 'the', 'playstation'), ('the', 'playstation', 'success'), ('playstation', 'succe
ss', 'led'), ('success', 'led', 'to'), ('led', 'to', 'a'), ('to', 'a', 'line'), ('a', 'line', 'of'), ('line', 'o
f', 'successors'), ('of', 'successors', 'beginning'), ('successors', 'beginning', 'with'), ('beginning', 'with',
'the'), ('with', 'the', 'playstation'), ('the', 'playstation', 'in'), ('playstation', 'in', 'in'), ('in', 'in',
'the'), ('in', 'the', 'same'), ('the', 'same', 'year'), ('same', 'year', 'sony'), ('year', 'sony', 'released'),
('sony', 'released', 'a'), ('released', 'a', 'smaller'), ('a', 'smaller', 'and'), ('smaller', 'and', 'cheaper'),
('and', 'cheaper', 'model'), ('cheaper', 'model', 'the'), ('model', 'the', 'ps'), ('the', 'ps', 'one')]]
```

ANSWER

Smaller k values are more sensitive in shingling because it requires less consecutive words to match. Larger k values make it less sensitive because more back to back words have to match.

Q3. Jaccard Similarity Computation

Compute the Jaccard similarity between all pairs of documents.

- Formula:

$$ J(A, B) = \frac{|A \cap B|}{|A \cup B|} $$

Interpretation

Measures overlap between shingle sets

- **$J(A, B)$** ranges from **0 to 1**.
- A value of **1** means the sets $A$ and $B$ are identical.
- A value of **0** means the sets $A$ and $B$ have no elements in common.

In [5]:

```python
# Jaccard Similarity Computation function
def jaccard(doc1_shingles, doc2_shingles):
    # Jaccard uses unqiue sets
    doc1_shingles = set(doc1_shingles)
    doc2_shingles = set(doc2_shingles)

    union = len(doc1_shingles | doc2_shingles)
    intersection = len(doc1_shingles & doc2_shingles)

    if intersection/union > 0:
        return intersection/union
    else:
        return 0
```

In [6]:

```python
# calculate jaccard between each pair of docs
for title1, doc1 in shingled_docs.items():
    for title2, doc2 in shingled_docs.items():
        if title1 <= title2:
            print(f"*{title1}* and *{title2}*: {round(jaccard(doc1,
                                                           doc2),4)}")
```

```
*iphone 7* and *iphone 7*: 1.0
*iphone 7* and *tesla cybertruck*: 0.0061
*iphone 7* and *tesla model s*: 0.0015
*iphone 7* and *playstation 5*: 0.0112
*iphone 7* and *playstation 4*: 0.0046
*tesla cybertruck* and *tesla cybertruck*: 1.0
*tesla cybertruck* and *tesla model s*: 0
*tesla model s* and *tesla model s*: 1.0
*playstation 5* and *tesla cybertruck*: 0.0025
*playstation 5* and *tesla model s*: 0.0017
*playstation 5* and *playstation 5*: 1.0
*iphone 6* and *iphone 7*: 0.0422
*iphone 6* and *tesla cybertruck*: 0
*iphone 6* and *tesla model s*: 0
*iphone 6* and *playstation 5*: 0.01
*iphone 6* and *iphone 6*: 1.0
*iphone 6* and *playstation 4*: 0.0041
*playstation 4* and *tesla cybertruck*: 0.0023
*playstation 4* and *tesla model s*: 0
*playstation 4* and *playstation 5*: 0.0399
*playstation 4* and *playstation 4*: 1.0
```

Q4. Thresholding for Near Duplicates

Choose a similarity threshold (e.g., 0.8 or 0.9).

- Identify which document pairs exceed this threshold.
- Justify your threshold choice.
- Flag near duplicates accordingly.

ANSWER

Our threshold is 0.5, because anything above this threshold means that the documents are more than 50% similar.

As we can see in our documents above, no document pairs exceed this threshold. This is probably because these wikipedia product descriptions have a lot of text, thus creating a lot of shingles. There are probably not many shingle matches in each document pairing due to that factor.

Another thing to consider is the word sensitivities, for example a shingle can be the same structurally, but the difference in "iphone 6" and "iphone 7" won't allow the intersection since the numbers are different.

The closest documents are "iphone 6" and "iphone 7" which have a score of 0.0422 (or 4.22%). This makes sense because these two iphone models are very similar.

Q5. Reflection

- What are the strengths and limitations of shingling for duplicate detection?
- How would you handle documents with minor formatting differences?
- Could shingling be extended to multimedia content?

Instructions:

- Use headings or bold labels
- Use markdown formatting to keep your notebook organized and readable.

ANSWER

Shingling is very good with being sensitive to major changes (depending on k), and it can be scaled to large documents since we are comparing k-sized lists in each document to retrieve a score.

It has a weakness of not being able to recognize sometimes semantically correct or differently phrased pairings. As mentioned above, the slight difference between "iphone6" and "iphone7" can lower the score significantly, even though they are very similar contextually.

Techniques such as synonym expansion may help with minor formatting differences, as we can expand the vocabulary but keep the context of the sentences the same. This approach would create a more accurate and advanced score compared to what we have now.

Shingling could most likely be used in multimedia content since the logic does not change of the algorithm. In images, we can break down pixels by k sized sets and find a similarity score of 2 pictures, same with audio trimmings.

Part 2: Detecting Web Spam in IR and Web Search

Q6. Keyword Stuffing Detection

Write a Python function that detects keyword stuffing in a document (web page).

- Accept a string of text and a target keyword.
- Count the number of times the keyword appears.
- Flag the document if the keyword frequency exceeds a threshold (e.g., 5).
- Normalize the text before counting.

Example Input:

`text = "Buy Maui golf real estate, Maui golf homes, Maui golf villas..."` `keyword = "Maui"`

Expected Output:

`True  # Keyword stuffing detected`

In [7]:

```
def keyword_stuff(text, keyword):
    keyword = keyword.lower()
    # Normalize and tokenize
    tokenized_text = nltk.word_tokenize(text)
    lowercased_text = [token.lower() for token in tokenized_text]
    normalized_text = [
        token for token in lowercased_text if token.isalpha() and token.isascii()
    ]
    occurence = 0
    for word in normalized_text:
        if word == keyword:
            occurence += 1
    if occurence >= 5:
        print("Keyword stuffing detected")
        return True
    else:
        return False
```

In [8]:

```
keyword_stuff("""Buy Maui golf real estate, Maui golf homes, Maui golf villas,
                Maui resorts,Maui soccer, Maui""","Maui")
```

```
Keyword stuffing detected
```

Out[8]:

```
True
```

In [9]:

```
keyword_stuff("Buy Maui golf real estate, Maui golf homes, Maui golf villas, Maui resorts","Maui")
```

Out[9]:

```
False
```

Q7. Hidden Text Detection

Simulate and detect hidden text in HTML using CSS-based concealment.

- Accept an HTML string.
- Search for patterns like `style="color:white"` or `display:none`.
- Return True if hidden text is detected.

Example Input:

`html = '<div style="color:white">Hidden keyword</div>'`

Expected Output:

`True`

In [10]:

```
def hidden_text(html):
    html = html.lower()

    hidden_text = [
        'display: none',
        'display:none',
        'color:white',
        'color: white',
        'color:#ffffff',
        'color: #ffffff',
        'font-size:0',
        'font-size: 0',
        'opacity:0',
        'opacity: 0'
    ]

    for text in hidden_text:
        if text in html:
            return True
    return False
```

In [11]:

```
hidden_text('<div style="color:white">Hidden keyword</div>')
```

Out[11]:

```
True
```

In [12]:

```
hidden_text('<div style="color:black">Hidden keyword</div>')
```

Out[12]:

```
False
```

Q8. Cloaking Detection

Compare crawler and user views of a webpage to detect cloaking.

- Accept two strings: one for crawler view, one for user view.
- Normalize both views.
- Return True if the views differ significantly.

Example Input:

`crawler_view = "Buy cheap flights to Rome" user_view = "Welcome to our travel blog!"`

Expected Output:

`True  # Cloaking detected`

In [13]:

```
def cloaking(crawler, user):
    crawler = crawler.lower()
    # Normalize and tokenize
    tokenized_crawler = nltk.word_tokenize(crawler)
    lowercased_crawler = [token.lower() for token in tokenized_crawler]
    normalized_crawler = [
        token for token in lowercased_crawler if token.isalpha() and token.isascii()
    ]

    user = user.lower()
    # Normalize and tokenize
    tokenized_user = nltk.word_tokenize(user)
    lowercased_user = [token.lower() for token in tokenized_user]
    normalized_user = [
        token for token in lowercased_user if token.isalpha() and token.isascii()
    ]

    # Shingling both user and crawler
    crawler_shingle = shingle(normalized_crawler)
    user_shingle = shingle(normalized_user)

    # Jaccard similarity
    similarity = jaccard(crawler_shingle, user_shingle)
    # Let's use threshold of 0.5 since 50% similarity is good
    if similarity < 0.5:
        print("Cloaking Detected")
        return True
    return False
```

In [14]:

```
cloaking("Buy cheap flights to Rome", "Welcome to our travel blog!")
```

```
Cloaking Detected
```

Out[14]:

```
True
```

In [15]:

```
cloaking("Buy cheap flights to Rome", "Buy cheap flights to Italy")
```

Out[15]:

```
False
```

Q9. Doorway Page Detection

Detect doorway pages that redirect users immediately.

- Accept an HTML string.
- Search for redirect patterns like `<meta http-equiv='refresh'>` or `window.location.href`.
- Return True if a redirect is found.

Example Input:

`html = "<meta http-equiv='refresh' content='0;url=https://ads.com'>"`

Expected Output:

`True`

In [16]:

```
def doorway_detect(html):
    html = html.lower()

    doorway_tags = [
        "http-equiv='refresh'",
        'http-equiv="refresh"',
        'http-equiv=refresh',
        "window.location.href"
    ]

    for text in doorway_tags:
        if text in html:
            return True
    return False
```

In [17]:

```
doorway_detect("<meta http-equiv='refresh' content='0;url=https://ads.com'>")
```

Out[17]:

```
True
```

In [18]:

```
doorway_detect('<div style="color:white">Hidden keyword</div>')
```

Out[18]:

```
False
```

Q10. Link Spam Detection

Detect link spam based on repeated or artificial inbound links.

- Accept a list of inbound URLs.
- Count duplicates and diversity.
- Return True if the number of unique links is low compared to total links.

Example Input:

`links = ["siteA.com", "siteA.com", "siteA.com", "siteB.com"]`

Expected Output:

`True # Link spam detected`

In [19]:

```
def link_spam(url_list):
    if len(url_list) == 0:
        return False

    # Lowercase first
    links = [url.lower() for url in url_list]

    total_urls = len(links)
    unique_urls = len(set(links))

    ratio = unique_urls / total_urls
    # using threshold of 0.5
    if ratio <= 0.5:
        print("Link spam detected")
        return True
    return False
```

In [20]:

```
link_spam(["siteA.com", "siteA.com", "siteA.com", "siteB.com"])
```

```
Link spam detected
```

Out[20]:

```
True
```

In [21]:

```
link_spam(["siteA.com", "siteB.com", "siteC.com", "siteD.com"])
```

Out[21]:

```
False
```

Q11: Estimate Content Quality of a Web Page

Objective:
Write a Python function that estimates the content quality of a web page based on its visible text. The function should return a score between 0 and 1, where higher values indicate better-written, more informative content.

Background:
Content quality is a key signal in search engine ranking. Pages with rich vocabulary, well-structured sentences, and minimal filler (e.g., stopwords) tend to be more useful to users and rank higher. This task simulates how search engines might heuristically assess page quality.

Instructions:

1. **Input:**
   - Accept a string containing raw HTML content of a web page.
2. **Text Extraction:**
   - Use `BeautifulSoup` to remove `<script>` and `<style>` tags.
   - Extract visible text and normalize whitespace.
3. **Tokenization and Preprocessing:**
   - Use `nltk.word_tokenize()` to split text into words.
   - Use `nltk.sent_tokenize()` to split text into sentences.
   - Filter out non-alphabetic tokens and lowercase all words.
4. **Heuristic Scoring:**
   Compute the following metrics:
   - **Average sentence length**: total words / total sentences
     (ideal range: 15–25 words per sentence)
   - **Vocabulary richness**: unique words / total words
     (higher is better)
   - **Stopword ratio**: proportion of stopwords in the text
     (lower is better)
5. **Final Score:**
   Combine the metrics using weighted scoring:
   - 40% weight for sentence length (normalized to 0–1)
   - 40% weight for vocabulary richness
   - 20% weight for inverse stopword ratio
6. **Output:**
   - Return a float score between 0 and 1, rounded to three decimal places.

Example Usage:

```
html_content = "<html><body><h1>Welcome</h1><p>Python is a versatile language used in many domains.
</p></body></html>"
score = estimate_webpage_content_quality(html_content)
print("Content Quality Score:", score)
```

In [22]:

```
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize

def estimate_webpage_content_quality(html):
    # Text extraction
    bs = BeautifulSoup(html, 'html.parser')

    for tag in bs(['script', 'style']):
        tag.decompose()
    text = bs.get_text(separator=' ', strip=True)
    text = ' '.join(text.split())

    if not text:
        return 0.0

    # Tokenization and preprocessingg
    # lowercasing
    sentences = nltk.sent_tokenize(text)
    text = text.lower()
    words = nltk.word_tokenize(text)

    # filter out non alphabetic tokens
    words = [
        word for word in words if word.isalpha()]

    if len(sentences)==0 or len(words)==0:
        return 0.0

    # scores
    avg_sen_len = len(words)/len(sentences)
    vocab_richness = len(set(words))/len(words)
    stop_words = set(stopwords.words('english'))

    stopword_ratio = sum(1 for word in words if word in stop_words)/len(words)

    # Normalize average sentence length
    if avg_sen_len < 15:
        sen_score = avg_sen_len/15
    # Ideal ratio is from 15-25
    elif avg_sen_len <= 25:
        sen_score = 1
    else:
        sen_score = max(0, 1-(avg_sen_len-25)/25)

    # final score (Use 1-stopword_ratio) for the inverse
    score = (0.4*sen_score)+(0.4*vocab_richness)+(0.2*(1-stopword_ratio))
    return round(score,3)
```

In [23]:

```
html="<html><body><h1>Welcome</h1><p>Python is a versatile language used in many domains.</p></body></html>"
score = estimate_webpage_content_quality(html)
print("Content Quality Score:", score)
```

```
Content Quality Score: 0.807
```

Q12: Calculate Keyword Density of a Web Page

Objective:
Write a Python function that calculates the keyword density of a web page. Keyword density is the ratio of the number of times a specific keyword appears to the total number of words in the page's visible text.

Background:
Keyword stuffing is a common spam tactic where keywords are repeated excessively to manipulate search rankings. Measuring keyword density helps identify whether a page uses keywords naturally or abusively.

Instructions:

1. **Input:**
    - Accept two inputs:
        - `html` (str): Raw HTML content of the web page.
        - `keyword` (str): The keyword to measure.
2. **Text Extraction:**
    - Use `BeautifulSoup` to remove `<script>` and `<style>` tags.
    - Extract visible text and normalize whitespace.
3. **Tokenization and Preprocessing:**
    - Use `nltk.word_tokenize()` to split text into words.
    - Filter out non-alphabetic tokens and lowercase all words.
4. **Keyword Density Calculation:**
    - Count the number of times the keyword appears.
    - Divide by the total number of words.
    - Return the result as a float rounded to three decimal places.

Example Usage:

```
html_content = "<html><body><p>Buy Maui golf real estate, Maui golf homes, Maui golf villas...</p>
</body></html>"
density = calculate_keyword_density(html_content, keyword="Maui")
print("Keyword Density:", density)
```

In [24]:

```python
def calculate_keyword_density(html, keyword):
    # Text extraction
    bs = BeautifulSoup(html, 'html.parser')

    for tag in bs(['script', 'style']):
        tag.decompose()

    text = bs.get_text(separator=' ', strip=True)
    text = ' '.join(text.split())

    if not text:
        return 0.0

    # Tokenization and preprocessingg
    # lowercasing
    text = text.lower()
    words = nltk.word_tokenize(text)

    # filter out non alphabetic tokens
    words = [
        word for word in words if word.isalpha()]

    if len(words)==0:
        return 0.0

    occurrence = 0.0
    for word in words:
        if word == keyword.lower():
            occurrence += 1

    ratio = occurrence / len(words)
    return round(ratio,3)
```

In [25]:

```python
html="<html><body><p>Buy Maui golf real estate, Maui golf homes, Maui golf villas...</p></body></html>"
density = calculate_keyword_density(html, keyword="Maui")
print("Keyword Density:", density)
```

```
Keyword Density: 0.273
```

Q13: Calculate Link Diversity of a Web Page

Objective:
Write a Python function that calculates the link diversity of a web page. Link diversity is the ratio of unique outbound links to total outbound links in the page's HTML.

Background:
Link spam involves creating many artificial or repeated links to falsely boost a page's authority. Measuring link diversity helps detect unnatural linking behavior.

Instructions:

1. **Input:**
    - Accept one input:
        - `html` (str): Raw HTML content of the web page.
2. **Link Extraction:**
    - Use `BeautifulSoup` to extract all `<a>` tags with `href` attributes.
    - Filter out internal links (e.g., those starting with `#`, `/`, or the same domain).
    - Normalize and collect all outbound links.
3. **Link Diversity Calculation:**
    - Count total outbound links.
    - Count unique outbound links.
    - Compute diversity as:
      $$ \text{Link Diversity} = \frac{\text{Unique Links}}{\text{Total Links}} $$
    - Return the result as a float rounded to three decimal places.

Link Diversity Metric

The Link Diversity metric is used to measure the variety of sources within a collection of links. It is calculated as the ratio of the number of unique links to the total number of links observed.

Interpretation

- **High Link Diversity (closer to 1):** Indicates a wide variety of sources, where most of the links point to different destinations.
- **Low Link Diversity (closer to 0):** Suggests that a few specific links are being repeated often, indicating a high concentration or reliance on a small number of sources. **Example Usage:**

```
html_content = '''
<html><body>
<a href="http://siteA.com">Link A</a>
<a href="http://siteA.com">Link A again</a>
<a href="http://siteB.com">Link B</a>
</body></html>
'''
diversity = calculate_link_diversity(html_content)
print("Link Diversity:", diversity)
```

In [26]:
```
def calculate_link_diversity(html):
    bs = BeautifulSoup(html, 'html.parser')
    urls = bs.find_all('a', href=True)

    # filter out inbound urls
    non_inB_urls = []
    for url in urls:
        href = url['href'].strip()

        if not href:
            continue
        if href.startswith('#') or href.startswith('/'):
            continue
        non_inB_urls.append(href.lower())

    if len(non_inB_urls)==0:
        return 0.0

    # diversity
    unique_links = len(set(non_inB_urls))
    total_links = len(non_inB_urls)

    diversity = unique_links/total_links
    return round(diversity,3)
```

In [27]:
```
html = '''
<html><body>
<a href="http://siteA.com">Link A</a>
<a href="http://siteA.com">Link A again</a>
<a href="http://siteB.com">Link B</a>
</body></html>
'''
diversity = calculate_link_diversity(html)
print("Link Diversity:", diversity)
```

```
Link Diversity: 0.667
```

Q14. Reflection and Extension

Write a structured summary of your findings for each of the questions from **Q6 to Q13**. For each technique or function you implemented, address the following aspects:

- **Weaknesses of the technique**
  Identify limitations, edge cases, or scenarios where the technique may fail or produce unreliable results.

- **Strengths of the technique**
  Highlight what the technique does well, including its efficiency, simplicity, or effectiveness in typical use cases.

- **Improvement opportunities**
  Explain how addressing the identified weaknesses could enhance the technique's accuracy, scalability, or robustness.

Instructions:

- Format your answers as separate paragraphs for each question (Q6 to Q13).
- Use headings or bold labels to clearly identify each question.
- Ensure your report includes meaningful insights and examples from your implementation.
- Use markdown formatting to keep your notebook organized and readable.

Q6

Keyword stuffing is a good technique to find inflated term frequencies given a keyword. This approach is very strong in considering which words are inflated given a threshold. There are some limitations to this however, such as long documents may be penalized due to their natural occurence of words that may appear a lot due to the fact that the document itself is large. An approach top fixing this weakness may be to implement a ratio of the term frequncy by the overall length of the document, and if this ratio exceeds a certain threshold, we can say that the keyword appears unnaturally in the document.

Q7

Finding out concealed keywords in css is useful due to the fact that we can find attributes that are invisible to users by parsed by crawlers. The algorithm is fast to execute and is good at catching common concelement techniques such as "display:none" and "opacity:0". A weakness this approach might have is that it can probably be evaded by creating more complex css strings, which would require more complex rules and domain knowledge by the developers trying to catch these techniques. One improvement may be to use regex patterns to account for similar hidden texts but with slight variations.

Q8

Cloaking detection is good for catching text which is different for the user and for the crawler, and using a similarity score to find how much they differ by. A weakness of this approach would be to access the crawler view which does not sound feasable in every scenario. Location based sites which change per region may be affected by the crawler view since it might not account for the different regions. One way to improve this algorithm might be to check html strings for more details on the document.

Q9

Doorway detection does a good job of finding common redirection patterns in html strings, it is fast in retrieval and efficient. There may be false positives in this scenario due to the algorithm catching very basic patterns, and a lot of legitimate sites may use redirects (such as mobile apps or portal logins). One improvement may be to only flag redirects which take 0 seconds, as that may be suspicious due to most notable websites giving a notice of redirect with a 5-second rate of redirection.

Q10

Link spam detection is good at identifying artifical link patterns where the same URL is repeadetly used. It can use a threshold parameter which is adjustable by the user as well. The drawback to link spam detection is that it cannot properly catch link spam accross many coordinated sites, link quality is also not taken into consideration, since a high quality webpage may be used multiple times in the current webpage. This algorithm can be heavily improved by combining it with the content quality algorithm of links that are in the html tag.

Q11

Content quality is good in terms that it uses ratios such as the average sentence length, stopwords, and unique vocabulary richness to estimate the quality of a page. The weakness this algorithm has is that it may be biased towards a certain type of document. Since the rules are very hard-coded, the content quality estimation may for example rank healthcare documents higher than sports documents due to the fact that it has more unique words, and less stopwords. One improvement for this will be to have an algorithm to estimate thresholds (such as average sentence length), for different document topics such as sports, healthcare, film, etc.

This algotihm cannot also detect factual accuracy of a document which may be an issue due to the fact that innacurate documents should automatically be lower in quality. One fix could be to add source credibility to have a better estimate.

Q12

Keyword density is good in terms of that it uses a ratio of term frequency by the length of the document to find inflated terms, rather than just the number of occurences by itself. A weakness is that it does not understand the semantic context of a document, and also it does not catch synonyms, or variations of a word (such a running, sprint, sprinting, and run are all similar). It can be improved using a thesauri and synonym expansion to find a more accurate score for the keyword density.

Q13

Link diversity of a webpage score is a good way of finding the link diversity of a webpage, by parsing the html tag and finding inflated links. It makes it easy to find pages that are repeats and spam urls. it cannot find the quality of the link however, and it may create false positives from verified links that have been cited multiple times in the webpage. We can improve this with our content estimation algorithm, by estimating the cited url's quality, and then flagging any occuring urls that have a low quality score.