Mini Project : Compilers

Data Summarization

Prepared by:
Ankur Chaudhari (41009)
Aditi Rupade (41048)
Abhilasha Talele (41054)

(Date)

Title :

Data Summarization

Problem Definition:

To summarize web article's data using POS tagging, chunking and Named-Entity Resolution.

Prerequisite:

Web Scraping, Compilers Phases, POS tagging, chunking, Named-Entity Resolution

Software Requirements:

Python 3.6 / Anaconda , Jupyter Notebook(optional), nltk package.

Outcomes:

Summarized data from data extracted through web articles.

Data Summarization :

Data Summarization is a simple term for a short conclusion of a big theory or a paragraph. This is something where you write the code and in the end, you declare the final result in the form of summarizing data. Data summarization has the great importance in the data mining. As nowadays a lot of programmers and developers work on big data theory. Earlier, you used to face difficulties to declare the result, but now there are so many relevant tools in the market where you can use in the programming or wherever you want in your data.

We are living in a digital world where data transfers in a second and it is much faster than a human capability. In the corporate field, employees work on a huge volume of data which is derived from different sources like Social Network, Media, Newspaper, Book, cloud media storage etc. But sometimes it may create difficulties for you to summarize the data. Sometimes you do not expect data volume because when you retrieve data from relational sources you can not predict that how much data will be stored in the database.

Approach

There are mainly two ways to make the summary. Extractive and Abstractive.

Extractive Method

Select relevant phrases of the input document and concatenate them to form a summary (like “copy-and-paste”).

- Pros: They are quite robust since they use existing natural-language phrases that are taken straight from the input.
- Cons: But they lack in flexibility since they cannot use novel words or connectors. They also cannot paraphrase like people sometimes do.

Abstractive Method

Generate a summary that keeps original intent. It’s just like humans do.

- Pros: They can use words that were not in the original input. It enables to make more fluent and natural summaries.
- Cons: But it is also a much harder problem as you now require the model to generate coherent phrases and connectors.

Implementation Using nltk Package :

In the first step, the data is extracted using web-scraping tool.

```
>>> scraped_data = urllib.request.urlopen('https://en.wikipedia.org/wiki/Hyundai')
```

The data is in html and xml format, hence it is parsed using the following function.

```
>>> parsed_article = bs.BeautifulSoup(article,'lxml')
```

Then, the data is preprocessed using regular expressions.

```
>>> article_text = re.sub(r'\[[0-9]*\]', ' ', article_text)
```

```
>>> article_text = re.sub(r'\s+', ' ', article_text)
```

This data is POS tagged using nltk package.

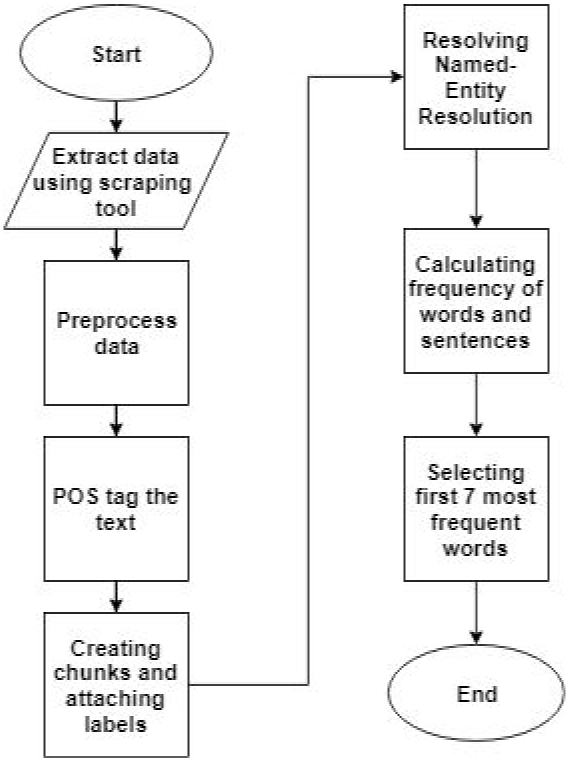
```
>>> tagged = nltk.pos_tag(text)
```

After tagging the data, various patterns of the following form are extracted from the data.

```
>>> pattern = 'NP: {<DT>?<JJ>*<NN>}'
```

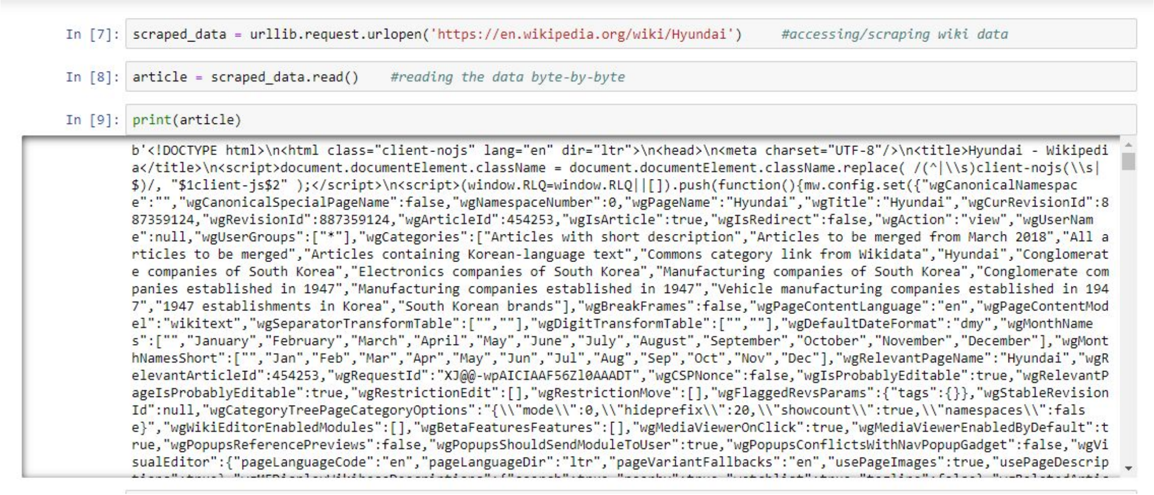
This data is then chunked and the tokens of the entities are found to resolve name-entity conflicts. The frequencies of words is calculated and total frequency of the sentence is calculated by summing the frequency of individual words. The first seven words are returned as summary of the article.

The flowchart of the above data processing techniques is as shown below :


Start
Extract data using scraping tool
Preprocess data
POS tag the text
Creating chunks and attaching labels
Resolving Named-Entity Resolution
Calculating frequency of words and sentences
Selecting first 7 most frequent words
End

Output :

The data after scraping from the web :



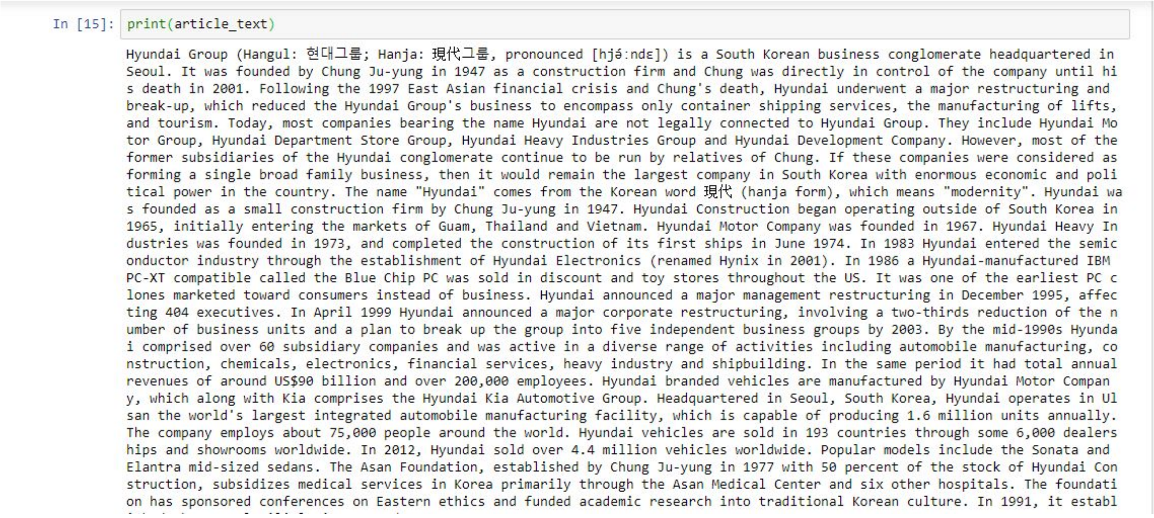
```
In [7]: scraped_data = urllib.request.urlopen('https://en.wikipedia.org/wiki/Hyundai')     #accessing/scraping wiki data

In [8]: article = scraped_data.read()    #reading the data byte-by-byte

In [9]: print(article)
```

```
b'<!DOCTYPE html>\n<html class="client-nojs" lang="en" dir="ltr">\n<head>\n<meta charset="UTF-8"/>\n<title>Hyundai - Wikipedi
a</title>\n<script>document.documentElement.className = document.documentElement.className.replace( /(^|\\s)client-nojs(\\s|
$)/, "$1client-js$2" );</script>\n<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespac
e":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"Hyundai","wgTitle":"Hyundai","wgCurRevisionId":8
87359124,"wgRevisionId":887359124,"wgArticleId":454253,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUserNam
e":null,"wgUserGroups":["*"],"wgCategories":["Articles with short description","Articles to be merged from March 2018","All a
rticles to be merged","Articles containing Korean-language text","Commons category link from Wikidata","Hyundai","Conglomerat
e companies of South Korea","Electronics companies of South Korea","Manufacturing companies of South Korea","Conglomerate com
panies established in 1947","Manufacturing companies established in 1947","Vehicle manufacturing companies established in 194
7","1947 establishments in Korea","South Korean brands"],"wgBreakFrames":false,"wgPageContentLanguage":"en","wgPageContentMod
el":"wikitext","wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthName
s":["","January","February","March","April","May","June","July","August","September","October","November","December"],"wgMont
hNamesShort":["","Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"],"wgRelevantPageName":"Hyundai","wgR
elevantArticleId":454253,"wgRequestId":"XJ@@-wpAICIAAF56Zl0AAADT","wgCSPNonce":false,"wgIsProbablyEditable":true,"wgRelevantP
ageIsProbablyEditable":true,"wgRestrictionEdit":[],"wgRestrictionMove":[],"wgFlaggedRevsParams":{"tags":{}},"wgStableRevision
Id":null,"wgCategoryTreePageCategoryOptions":"{\\"mode\\":0,\\"hideprefix\\":20,\\"showcount\\":true,\\"namespaces\\":fals
e}","wgWikiEditorEnabledModules":[],"wgBetaFeaturesFeatures":[],"wgMediaViewerOnClick":true,"wgMediaViewerEnabledByDefault":t
rue,"wgPopupsReferencePreviews":false,"wgPopupsShouldSendModuleToUser":true,"wgPopupsConflictsWithNavPopupGadget":false,"wgVi
sualEditor":{"pageLanguageCode":"en","pageLanguageDir":"ltr","pageVariantFallbacks":"en","usePageImages":true,"usePageDescrip
```

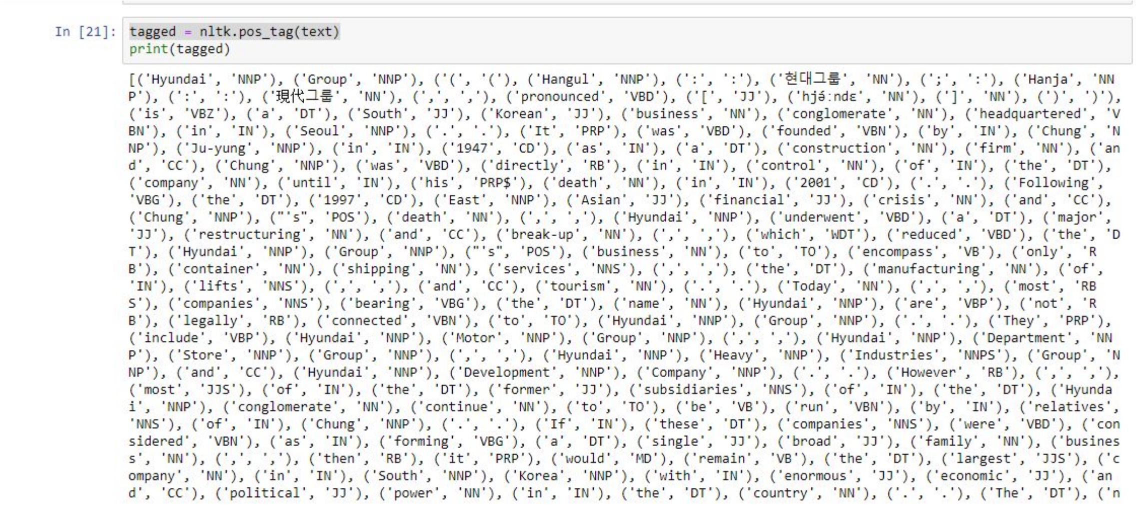
The data after preprocessing :



```
In [15]: print(article_text)

Hyundai Group (Hangul: 현대그룹; Hanja: 現代그룹, pronounced [hjə́ːndɛ]) is a South Korean business conglomerate headquartered in
Seoul. It was founded by Chung Ju-yung in 1947 as a construction firm and Chung was directly in control of the company until hi
s death in 2001. Following the 1997 East Asian financial crisis and Chung's death, Hyundai underwent a major restructuring and
break-up, which reduced the Hyundai Group's business to encompass only container shipping services, the manufacturing of lifts,
and tourism. Today, most companies bearing the name Hyundai are not legally connected to Hyundai Group. They include Hyundai Mo
tor Group, Hyundai Department Store Group, Hyundai Heavy Industries Group and Hyundai Development Company. However, most of the
former subsidiaries of the Hyundai conglomerate continue to be run by relatives of Chung. If these companies were considered as
forming a single broad family business, then it would remain the largest company in South Korea with enormous economic and poli
tical power in the country. The name "Hyundai" comes from the Korean word 現代 (hanja form), which means "modernity". Hyundai wa
s founded as a small construction firm by Chung Ju-yung in 1947. Hyundai Construction began operating outside of South Korea in
1965, initially entering the markets of Guam, Thailand and Vietnam. Hyundai Motor Company was founded in 1967. Hyundai Heavy In
dustries was founded in 1973, and completed the construction of its first ships in June 1974. In 1983 Hyundai entered the semic
onductor industry through the establishment of Hyundai Electronics (renamed Hynix in 2001). In 1986 a Hyundai-manufactured IBM
PC-XT compatible called the Blue Chip PC was sold in discount and toy stores throughout the US. It was one of the earliest PC c
lones marketed toward consumers instead of business. Hyundai announced a major management restructuring in December 1995, affec
ting 404 executives. In April 1999 Hyundai announced a major corporate restructuring, involving a two-thirds reduction of the n
umber of business units and a plan to break up the group into five independent business groups by 2003. By the mid-1990s Hyunda
i comprised over 60 subsidiary companies and was active in a diverse range of activities including automobile manufacturing, co
nstruction, chemicals, electronics, financial services, heavy industry and shipbuilding. In the same period it had total annual
revenues of around US$90 billion and over 200,000 employees. Hyundai branded vehicles are manufactured by Hyundai Motor Compan
y, which along with Kia comprises the Hyundai Kia Automotive Group. Headquartered in Seoul, South Korea, Hyundai operates in Ul
san the world's largest integrated automobile manufacturing facility, which is capable of producing 1.6 million units annually.
The company employs about 75,000 people around the world. Hyundai vehicles are sold in 193 countries through some 6,000 dealers
hips and showrooms worldwide. In 2012, Hyundai sold over 4.4 million vehicles worldwide. Popular models include the Sonata and
Elantra mid-sized sedans. The Asan Foundation, established by Chung Ju-yung in 1977 with 50 percent of the stock of Hyundai Con
struction, subsidizes medical services in Korea primarily through the Asan Medical Center and six other hospitals. The foundati
on has sponsored conferences on Eastern ethics and funded academic research into traditional Korean culture. In 1991, it establ
```

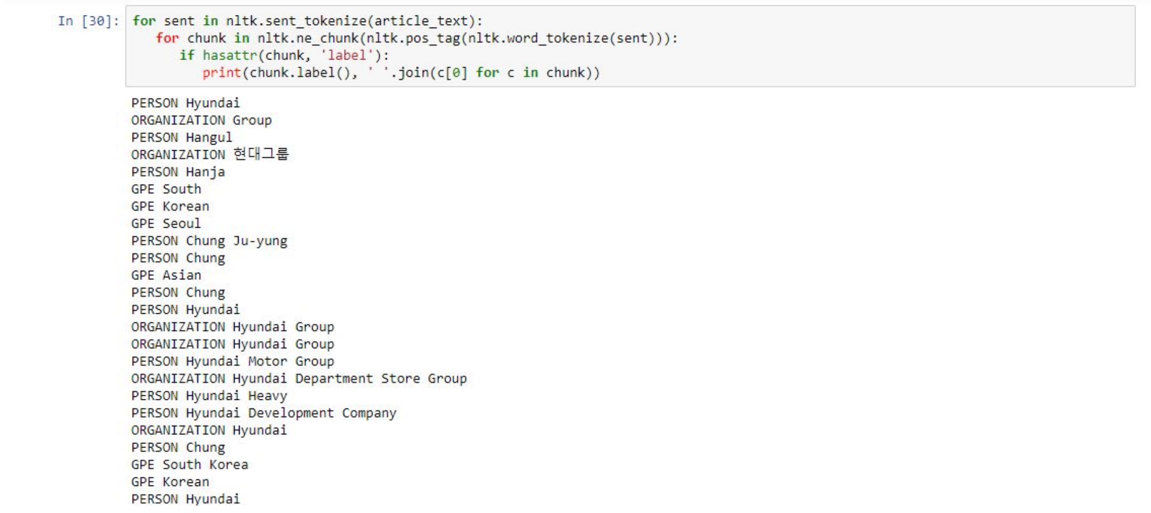
The data after POS tagging :



```
In [21]: tagged = nltk.pos_tag(text)
         print(tagged)

[('Hyundai', 'NNP'), ('Group', 'NNP'), ('(', '('), ('Hangul', 'NNP'), (':', ':'), ('현대그룹', 'NN'), (';', ':'), ('Hanja', 'NN
P'), (':', ':'), ('現代그룹', 'NN'), (',', ','), ('pronounced', 'VBD'), ('[', 'JJ'), ('hjə́ːndɛ', 'NN'), (']', 'NN'), (')', ')'),
('is', 'VBZ'), ('a', 'DT'), ('South', 'JJ'), ('Korean', 'JJ'), ('business', 'NN'), ('conglomerate', 'NN'), ('headquartered', 'V
BN'), ('in', 'IN'), ('Seoul', 'NNP'), ('.', '.'), ('It', 'PRP'), ('was', 'VBD'), ('founded', 'VBN'), ('by', 'IN'), ('Chung', 'N
NP'), ('Ju-yung', 'NNP'), ('in', 'IN'), ('1947', 'CD'), ('as', 'IN'), ('a', 'DT'), ('construction', 'NN'), ('firm', 'NN'), ('an
d', 'CC'), ('Chung', 'NNP'), ('was', 'VBD'), ('directly', 'RB'), ('in', 'IN'), ('control', 'NN'), ('of', 'IN'), ('the', 'DT'),
('company', 'NN'), ('until', 'IN'), ('his', 'PRP$'), ('death', 'NN'), ('in', 'IN'), ('2001', 'CD'), ('.', '.'), ('Following',
'VBG'), ('the', 'DT'), ('1997', 'CD'), ('East', 'NNP'), ('Asian', 'JJ'), ('financial', 'JJ'), ('crisis', 'NN'), ('and', 'CC'),
('Chung', 'NNP'), ("'s", 'POS'), ('death', 'NN'), (',', ','), ('Hyundai', 'NNP'), ('underwent', 'VBD'), ('a', 'DT'), ('major',
'JJ'), ('restructuring', 'NN'), ('and', 'CC'), ('break-up', 'NN'), (',', ','), ('which', 'WDT'), ('reduced', 'VBD'), ('the', 'D
T'), ('Hyundai', 'NNP'), ('Group', 'NNP'), ("'s", 'POS'), ('business', 'NN'), ('to', 'TO'), ('encompass', 'VB'), ('only', 'R
B'), ('container', 'NN'), ('shipping', 'NN'), ('services', 'NNS'), (',', ','), ('the', 'DT'), ('manufacturing', 'NN'), ('of',
'IN'), ('lifts', 'NNS'), (',', ','), ('and', 'CC'), ('tourism', 'NN'), ('.', '.'), ('Today', 'NN'), (',', ','), ('most', 'RB
S'), ('companies', 'NNS'), ('bearing', 'VBG'), ('the', 'DT'), ('name', 'NN'), ('Hyundai', 'NNP'), ('are', 'VBP'), ('not', 'R
B'), ('legally', 'RB'), ('connected', 'VBN'), ('to', 'TO'), ('Hyundai', 'NNP'), ('Group', 'NNP'), ('.', '.'), ('They', 'PRP'),
('include', 'VBP'), ('Hyundai', 'NNP'), ('Motor', 'NNP'), ('Group', 'NNP'), (',', ','), ('Hyundai', 'NNP'), ('Department', 'NN
P'), ('Store', 'NNP'), ('Group', 'NNP'), (',', ','), ('Hyundai', 'NNP'), ('Heavy', 'NNP'), ('Industries', 'NNPS'), ('Group', 'N
NP'), ('and', 'CC'), ('Hyundai', 'NNP'), ('Development', 'NNP'), ('Company', 'NNP'), ('.', '.'), ('However', 'RB'), (',', ','),
('most', 'JJS'), ('of', 'IN'), ('the', 'DT'), ('former', 'JJ'), ('subsidiaries', 'NNS'), ('of', 'IN'), ('the', 'DT'), ('Hyunda
i', 'NNP'), ('conglomerate', 'NN'), ('continue', 'NN'), ('to', 'TO'), ('be', 'VB'), ('run', 'VBN'), ('by', 'IN'), ('relatives',
'NNS'), ('of', 'IN'), ('Chung', 'NNP'), ('.', '.'), ('If', 'IN'), ('these', 'DT'), ('companies', 'NNS'), ('were', 'VBD'), ('con
sidered', 'VBN'), ('as', 'IN'), ('forming', 'VBG'), ('a', 'DT'), ('single', 'JJ'), ('broad', 'JJ'), ('family', 'NN'), ('busines
s', 'NN'), (',', ','), ('then', 'RB'), ('it', 'PRP'), ('would', 'MD'), ('remain', 'VB'), ('the', 'DT'), ('largest', 'JJS'), ('c
ompany', 'NN'), ('in', 'IN'), ('South', 'NNP'), ('Korea', 'NNP'), ('with', 'IN'), ('enormous', 'JJ'), ('economic', 'JJ'), ('an
d', 'CC'), ('political', 'JJ'), ('power', 'NN'), ('in', 'IN'), ('the', 'DT'), ('country', 'NN'), ('.', '.'), ('The', 'DT'), ('n
```

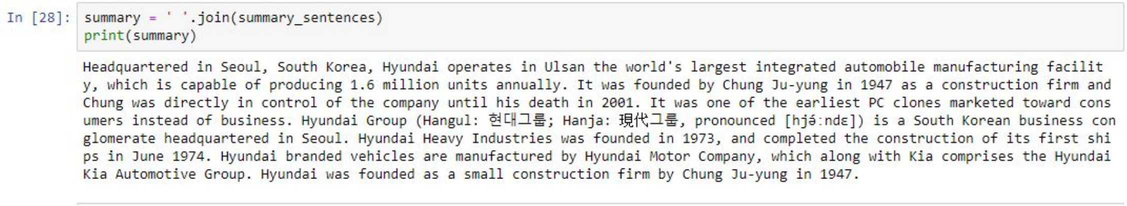
The data after Named-Entity Resolution :



```
In [30]: for sent in nltk.sent_tokenize(article_text):
            for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(sent))):
               if hasattr(chunk, 'label'):
                  print(chunk.label(), ' '.join(c[0] for c in chunk))
```

```
PERSON Hyundai
ORGANIZATION Group
PERSON Hangul
ORGANIZATION 현대그룹
PERSON Hanja
GPE South
GPE Korean
GPE Seoul
PERSON Chung Ju-yung
PERSON Chung
GPE Asian
PERSON Chung
PERSON Hyundai
ORGANIZATION Hyundai Group
ORGANIZATION Hyundai Group
PERSON Hyundai Motor Group
ORGANIZATION Hyundai Department Store Group
PERSON Hyundai Heavy
PERSON Hyundai Development Company
ORGANIZATION Hyundai
PERSON Chung
GPE South Korea
GPE Korean
PERSON Hyundai
```

Summarized Data :



```
In [28]: summary = ' '.join(summary_sentences)
         print(summary)
```

```
Headquartered in Seoul, South Korea, Hyundai operates in Ulsan the world's largest integrated automobile manufacturing facilit
y, which is capable of producing 1.6 million units annually. It was founded by Chung Ju-yung in 1947 as a construction firm and
Chung was directly in control of the company until his death in 2001. It was one of the earliest PC clones marketed toward cons
umers instead of business. Hyundai Group (Hangul: 현대그룹; Hanja: 現代그룹, pronounced [hjə́ːndɛ]) is a South Korean business con
glomerate headquartered in Seoul. Hyundai Heavy Industries was founded in 1973, and completed the construction of its first shi
ps in June 1974. Hyundai branded vehicles are manufactured by Hyundai Motor Company, which along with Kia comprises the Hyundai
Kia Automotive Group. Hyundai was founded as a small construction firm by Chung Ju-yung in 1947.
```

Approved by :

Prof. Yogita Narwadkar