1. Data Pre-Processing

February 9, 2023

[1]:

```
text = "2022 came to a celebratory end for BTS, especially member J-Hope, who␣
 ↪was in New York, making another fabulous display of his skills  While away␣
 ↪from the members, he seemed to have enjoyed it to the fullest with a solo␣
 ↪performance at Dick Clark&rsquo;s New Year&rsquo;s Rockin&rsquo; Eve where␣
 ↪he did a live stage of 3 songs, solo track &lsquo;= (Equal Sign)&rsquo;,␣
 ↪&lsquo;Chicken Noodle Soup&rsquo; his collaboration track with Becky G, and␣
 ↪BTS&rsquo; &lsquo;Butter&rsquo; (Holiday Remix)  He officially became only␣
 ↪the second South Korean soloist to perform at the event, following PSY  This␣
 ↪was also J-Hope&rsquo;s third time at Dick Clark&rsquo;s New Year&rsquo;s␣
 ↪Rockin&rsquo; Eve after group stages with BTS in 2017 and 2019␣
 ↪\n\nJ-Hope&rsquo;s live\n\nAfter returning from Times Square where he␣
 ↪performed as the penultimate act alongside multiple other singers from␣
 ↪around the world and spoke to Ryan Seacrest, the host of the show, J-Hope␣
 ↪was back at his hotel room and turned on a live broadcast to speak with his␣
 ↪fans  As they congratulated him on his sparkling performance, being the␣
 ↪perfectionist that he is, J-Hope expressed his sadness about being unable to␣
 ↪use his voice to the fullest and talked about the slipping incident due to␣
 ↪the rainy weather during his rehearsal stage  \n\nJ-Hope on other BTS␣
 ↪members&rsquo; New Year wishes\n\nAs soon as it turned 12, and the New Year␣
 ↪began in South Korea, BTS members Jungkook, Jimin, and V took to the fan␣
 ↪community platform Weverse to share their wishes with the fans as well as␣
 ↪discuss their plans for the coming year  Jungkook and V kept it brief,␣
 ↪wishing for a successful and happy year ahead while Jimin wrote a big letter␣
 ↪to his fans about the feelings he had seemingly bottled for a long time␣
 ↪about his wishes to release new music soon for which he has been meeting up␣
 ↪with composers \n\nThousands of miles away, member J-Hope was his cheeky␣
 ↪self as he unleashed all his love for his fellow BTS members  He began␣
 ↪commenting &lsquo;love you&rsquo; on their posts with cute words and␣
 ↪received laughs from them in return \n\nJin&rsquo;s call to J-Hope\n\nOn␣
 ↪being asked about BTS&rsquo; oldest member Jin who became the first from the␣
```

[2]:
```
import re
import string
import contractions
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
```

[3]:
```
class Data_Cleaning:
    def __init__(self) -> None:
        self.__lxmlParser = "lxml"                          # lxml parser
        self.__nextLine = "(\\n)+"                          # re for new line␣
 ↪characters
        self.__tabs = "(\\t)+"                              # re for tabs
        self.__invalidSpaces = "(  )+"                      # re for more than 2␣
 ↪spaces between words
        self.__punctuations = "[,‘’():=!"?"]"               # re for punctuations
        self.__contractions = contractions                  # to get common␣
 ↪contractions
        self.__stopwords = stopwords.words("english")       # gives us a list of␣
 ↪stopwords in english

    # to remove HTML Tags and Entites
    def __removing_html_tags_and_entities(self, data: str) -> str:
        return BeautifulSoup(data, self.__lxmlParser).get_text(strip=True)

    # to add fullstops at appropriate positions and removing more than 2␣
 ↪consecutive fullstops
    def __modification_full_stops(self, data: str) -> str:
        __res = ""
        data =  re.sub(self.__invalidSpaces, '.', \
                re.sub(self.__tabs, '', \
                re.sub(self.__nextLine, '.', data)))
        for i in data.split("."):
            if len(i.strip()) != 0:
                __res += i.strip() + ". "
        return __res

    # to convert contractions to expanded words
    def __removing_contractions(self, data: str) -> str:
        __res = ""
        for i in data.split(" "):
            __res += self.__contractions.fix(i) + " "
        return __res

    # to remove punctuations
    def __removing_punctuations(self, data: str) -> str:
        data = data.lower()
        return re.sub(self.__punctuations, '', data)
```

```
    # to remove stopwords?"
    def __removing_stopwords(self, data: str) -> str:
        __res = ""
        for i in data.split(" "):
            if i not in self.__stopwords:
                __res += i + " "
        return __res


    # the actual function that does the work
    def pre_processor(self, data: str) -> str:
        __lvl1_preprocessing = self.__removing_contractions( \
                            self.__modification_full_stops( \
                            self.__removing_html_tags_and_entities(data)))

        __lvl2_preprocessing = self.__removing_punctuations( \
                            self.__removing_stopwords(__lvl1_preprocessing))
        return [__lvl1_preprocessing, __lvl2_preprocessing]
```

[4]:
```
p = Data_Cleaning()
cleaned_data, structured_data = p.pre_processor(text)
print(f"Actual Data:\n{cleaned_data}\n")
print(f"Cleaned Data:\n{structured_data}\n")
```

```
Actual Data:
2022 came to a celebratory end for BTS, especially member J-Hope, who was in New
York, making another fabulous display of his skills. While away from the
members, he seemed to have enjoyed it to the fullest with a solo performance at
Dick Clark's New Year's Rockin' Eve where he did a live stage of 3 songs, solo
track '= (Equal Sign)', 'Chicken Noodle Soup' his collaboration track with Becky
G, and BTS' 'Butter' (Holiday Remix). He officially became only the second South
Korean soloist to perform at the event, following PSY. This was also J-Hope's
third time at Dick Clark's New Year's Rockin' Eve after group stages with BTS in
2017 and 2019. J-Hope's live. After returning from Times Square where he
performed as the penultimate act alongside multiple other singers from around
the world and spoke to Ryan Seacrest, the host of the show, J-Hope was back at
his hotel room and turned on a live broadcast to speak with his fans. As they
congratulated him on his sparkling performance, being the perfectionist that he
is, J-Hope expressed his sadness about being unable to use his voice to the
fullest and talked about the slipping incident due to the rainy weather during
his rehearsal stage. J-Hope on other BTS members' New Year wishes. As soon as it
turned 12, and the New Year began in South Korea, BTS members Jungkook, Jimin,
and V took to the fan community platform Weverse to share their wishes with the
fans as well as discuss their plans for the coming year. Jungkook and V kept it
brief, wishing for a successful and happy year ahead while Jimin wrote a big
letter to his fans about the feelings he had seemingly bottled for a long time
about his wishes to release new music soon for which he has been meeting up with
```

composers. Thousands of miles away, member J-Hope was his cheeky self as he unleashed all his love for his fellow BTS members. He began commenting ‘love you’ on their posts with cute words and received laughs from them in return. Jin’s call to J-Hope. On being asked about BTS’ oldest member Jin who became the first from the group to enlist in the military on December 13, J-Hope recalled how he was called by the member with a different phone number and almost missed it. J-Hope said, "Right before I was about to sleep on the 31st, I got a call from Jin and asked him how he was doing to which he said J-Hope, pick up your phone. I told him I did not know this number, how would I know it was his?". The ‘Arson’ hitmaker spoke with a smile on his face about how he felt happy hearing Jin’s voice. It comforted him and J-Hope mentioned how he remembered all those moments he spent with Jin. He assured the fans by saying that Jin seemed to be healthy and doing well in the military. So in place of Jin, he shared that he was well and asked the fans to not worry. The lovely duo, BTS’ Jin and J-Hope nicknamed 2seok have always lightened the fans’ hearts with this interaction. https://twitter com/nightstar1201/status/1609428759454822401Jimin X Taeyang. BTS member Jimin and BIGBANG member Taeyang are a collaboration nobody would have expected. However, in December 2022, it was reported by industry officials that the two are working together on a release that will soon be revealed to the world. To this, Taeyang’s then-agency, YG Entertainment, which he has since departed, replied by saying that they cannot confirm anything at the moment and asked fans to anticipate Taeyang’s activities. On January 1, Taeyang who is currently a free agent, not having signed with anyone for his solo activities as he continues to be with YG Entertainment for any content related to BIGBANG, shared a rare update on his Instagram account. 2 photos were shared with a caption of the hashtag #2023. In black and white, 2 people could be seen in the photos, their backs to the camera. While one could easily be spotted as Taeyang himself, the other one was not tagged. Fans went into action and soon began investigating every detail of the photo. The other person seemed to be Jimin, as his hands, fluffy hair, and earrings seem to match the BTS member. Though no official announcement was made from either side, it seems that the singers have decided to give a green signal from their ends. Taeyang is rumoured to have been preparing for a January 2023 comeback so it seems as though we can expect the reports any time now. Once confirmed, this could very well be the biggest release of the year and one of the most hyped collaborations in K-pop history!. Meanwhile, BIGBANG’s G-Dragon also announced that he is working on an album and hopes to release it in 2023. Stay updated with the latest Hallyu news on: Instagram, YouTube, Twitter, Facebook and Snapchat. ALSO READ: BTS’ J-Hope, TXT wow at Dick Clark’s New Year’s Rockin’ Eve: 5 highlights from their stages.

Cleaned Data:
2022 came celebratory end bts especially member j-hope new york making another fabulous display skills. while away members seemed enjoyed fullest solo performance dick clarks new years rockin eve live stage 3 songs solo track equal sign chicken noodle soup collaboration track becky g bts butter holiday remix. he officially became second south korean soloist perform event following psy. this also j-hopes third time dick clarks new years rockin eve group stages bts 2017 2019. j-hopes live. after returning times square performed penultimate

act alongside multiple singers around world spoke ryan seacrest host show j-hope back hotel room turned live broadcast speak fans. as congratulated sparkling performance perfectionist is j-hope expressed sadness unable use voice fullest talked slipping incident due rainy weather rehearsal stage. j-hope bts members new year wishes. as soon turned 12 new year began south korea bts members jungkook jimin v took fan community platform weverse share wishes fans well discuss plans coming year. jungkook v kept brief wishing successful happy year ahead jimin wrote big letter fans feelings seemingly bottled long time wishes release new music soon meeting composers. thousands miles away member j-hope cheeky self unleashed love fellow bts members. he began commenting love you posts cute words received laughs return. jins call j-hope. on asked bts oldest member jin became first group enlist military december 13 j-hope recalled called member different phone number almost missed it. j-hope said right i sleep 31st i got call jin asked said j-hope pick phone. i told i know number would i know his. the arson hitmaker spoke smile face felt happy hearing jins voice. it comforted j-hope mentioned remembered moments spent jin. he assured fans saying jin seemed healthy well military. so place jin shared well asked fans worry. the lovely duo bts jin j-hope nicknamed 2seok always lightened fans hearts interaction. https//twitter com/nightstar1201/status/1609428759454822401jimin x taeyang. bts member jimin bigbang member taeyang collaboration nobody would expected. however december 2022 reported industry officials two working together release soon revealed world. to this taeyangs then-agency yg entertainment since departed replied saying cannot confirm anything moment asked fans anticipate taeyangs activities. on january 1 taeyang currently free agent signed anyone solo activities continues yg entertainment content related bigbang shared rare update instagram account. 2 photos shared caption hashtag #2023. in black white 2 people could seen photos backs camera. while one could easily spotted taeyang himself one tagged. fans went action soon began investigating every detail photo. the person seemed jimin hands fluffy hair earrings seem match bts member. though official announcement made either side seems singers decided give green signal ends. taeyang rumoured preparing january 2023 comeback seems though expect reports time now. once confirmed could well biggest release year one hyped collaborations k-pop history. meanwhile bigbangs g-dragon also announced working album hopes release 2023. stay updated latest hallyu news on instagram youtube twitter facebook snapchat. also read bts j-hope txt wow dick clarks new years rockin eve 5 highlights stages.