In [4]:
```
import re
from bs4 import BeautifulSoup
import os
import pandas as pd
import requests
```

In [23]:
```
file_path = '/Users/meredithfan/Desktop/0001017386-21-000108.txt'
with open(file_path, 'r', encoding= 'utf-8') as file:
    file_content = file.read()
```

In [24]:
```
pattern = re.compile(r'\bDiscussion\s+and\s+Analysis\s+of\s+Financi
matches = re.finditer(pattern, file_content)
indices = [match.start() for match in matches]

if len(indices) >= 1:
    start_index_1 = indices[0]
    if len(indices) >= 2:
        start_index_2 = indices[1]
    else:
        start_index_2 = -1
else:
    # If no matches found for the plain text pattern, check the bol
    pattern_html = re.compile(r'<b>\s*Management\s+Discussion\s+and
    matches_html = re.finditer(pattern_html, file_content)
    indices_html = [match.start() for match in matches_html]

    if len(indices_html) >= 1:
        start_index_1 = indices_html[0]
        if len(indices_html) >= 2:
            start_index_2 = indices_html[1]
        else:
            start_index_2 = -1
    else:
        # Handle the case when no occurrence is found for both patt
        start_index_1 = -1
        start_index_2 = -1
```

In [25]:
```
attern = re.compile(r"Disclosure[s]? About Market Risk", re.IGNORECA
atches = re.finditer(pattern, file_content)
ndices = [match.start() for match in matches]

f len(indices) >= 1:
   end_index_1 = indices[0]
   if len(indices) >= 2:
       end_index_2 = indices[1]
   else:
       end_index_2 = -1
lse:
   end_index_1 = -1
   end_index_2 = -1

f end_index_1 == -1:
   pattern = re.compile(r"Control[s]? and Procedure[s]?", re.IGNOREC
   matches = re.finditer(pattern, file_content)
   indices = [match.start() for match in matches]

   if len(indices) >= 1:
       end_index_1 = indices[0]
       if len(indices) >= 2:
           end_index_2 = indices[1]
       else:
           end_index_2 = -1
   else:
       end_index_1 = -1
       end_index_2 = -1

f end_index_2==-1 and start_index_2==-1 and start_index_1<end_index_
   extracted_text = file_content[start_index_1:end_index_1]
lif start_index_2<end_index_2:
   extracted_text = file_content[start_index_2:end_index_2]
lse:
   extracted_text=""
f start_index_1==-1:
   extracted_text=''

tml_text = extracted_text
oup = BeautifulSoup(html_text, 'html.parser')
ables = soup.find_all('table')
or table in tables:
   table.decompose()

odified_html_text = str(soup)

oup = BeautifulSoup(modified_html_text, 'html.parser')
ext = soup.get_text()
```

In [26]:
```
string = text
clean_string = ''.join(string.replace('\n', ' . ').replace('\u200b'

clean_string = re.sub(r"(?i)table of contents", "", clean_string)
clean_string = re.sub(r"\bquantitative and Qualitative\b", "", clea

clean_string = re.sub(r"\s*(?:item\s*7a\.?|item\s*4\.?|item\s*3\.?|
clean_string = "Management's " + clean_string

clean_string =  re.sub(r"[•;]", " . ", clean_string)
clean_string =  re.sub(r'\s*\.\s*(\.\s*)*', '. ', clean_string)
clean_string = clean_string.strip()
```

In [27]:
```
print(clean_string)
```

```
Management's
```

In [28]:
```
def count_words(input_string):
    # Remove leading and trailing whitespaces (optional)
    input_string = input_string.strip()

    # Split the string into words based on spaces (you can use other
    words_list = input_string.split()

    # Count the number of words in the list
    word_count = len(words_list)

    return word_count
```

In [29]:
```
num_words= count_words(clean_string)
num_words
```

Out[29]:
```
1
```

In [ ]: