Tokenization

- Tokenization is the process of splitting input sequence into TOKENS. The words are an example of tokens in a sentence.

In [1]:

```
import pandas as pd
import numpy as np
import re
import matplotlib.pyplot as plt
```

- A word is a meaningful sequence of characters.
- Word boundaries in English are spaces or punctuation.
- In German there are compound words which are written without spaces
  - “Rechtsschutzversicherungsgesellschaften” stands for “insurance companies which provide legal protection”
- In Japanese there are no spaces at all!
  - Butyoucanstillreaditright?

Tokenization is a process that splits input sequence into tokens

- Token as useful unit for semantic processing, they can be a word, sentence, paragraph, etc.

In [2]:

```
test_seq = "Andrew Ng is datajango's guru, isn't it. Anderw is a professor in Stanford Univ
```

In [3]:

```
from nltk.tokenize import WordPunctTokenizer
wp_tokenizer = WordPunctTokenizer()
tokens = wp_tokenizer.tokenize(test_seq)
print(tokens)
```

```
['Andrew', 'Ng', 'is', 'datajango', "'", 's', 'guru', ',', 'isn', "'", 't',
'it', '.', 'Anderw', 'is', 'a', 'professor', 'in', 'Stanford', 'University',
',', 'San', 'Francisco', ',', 'California', '.']
```

In [4]:

```
from nltk.tokenize import TreebankWordTokenizer
tb_tokenizer = TreebankWordTokenizer()
tokens = tb_tokenizer.tokenize(test_seq)
print(tokens)
```

```
['Andrew', 'Ng', 'is', 'datajango', "'s", 'guru', ',', 'is', "n't", 'it.',
'Anderw', 'is', 'a', 'professor', 'in', 'Stanford', 'University', ',', 'Sa
n', 'Francisco', ',', 'California', '.']
```

In [5]:

```
len(tokens)
```

Out[5]:

```
23
```

In [6]:

```
from nltk.tokenize import WhitespaceTokenizer
whp_tokenizer = WhitespaceTokenizer()
tokens = whp_tokenizer.tokenize(test_seq)
print(tokens)
```

```
['Andrew', 'Ng', 'is', "datajango's", 'guru,', "isn't", 'it.', 'Anderw', 'i
s', 'a', 'professor', 'in', 'Stanford', 'University,', 'San', 'Francisco,',
'California.']
```

Stanford Core NLP

- https://github.com/nltk/nltk/wiki/Stanford-CoreNLP-API-in-NLTK (https://github.com/nltk/nltk/wiki/Stanford-CoreNLP-API-in-NLTK)
- https://stanfordnlp.github.io/CoreNLP/ (https://stanfordnlp.github.io/CoreNLP/)
- Stanford CoreNLP is written in Java; recent releases require Java 1.8+ (https://www.oracle.com/technetwork/java/javaee/downloads/jdk8-downloads-2133151.html (https://www.oracle.com/technetwork/java/javaee/downloads/jdk8-downloads-2133151.html)). You need to have Java installed to run CoreNLP. However, you can interact with CoreNLP via the command-line or its web service; many people use CoreNLP while writing their own code in Javascript, Python, or some other language.
  - goto stanford-corenlp-full
  - java -mx4g -cp "*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer -preload tokenize,ssplit,pos,lemma,ner,parse,depparse -status_port 9000 -port 9000 -timeout 15000 &

In [7]:

```
# from nltk.tokenize.stanford import StanfordTokenizer depricated
from nltk.parse.corenlp import CoreNLPParser
st_parser = CoreNLPParser(url='http://localhost:9000')
#st_parser = CoreNLPParser(url='http://[::]:9000/')
tokens = st_parser.tokenize(test_seq)
print(tokens)
```

```
<generator object GenericCoreNLPParser.tokenize at 0x00000213A2D83040>
```

Generator: "yield" key word makes a method generator

In [8]:

```
def squares(nums):
    sqr = []
    for n in nums:
        sqr.append(n**2)
    return sqr

print(squares([1,2,3,4]))
```

```
[1, 4, 9, 16]
```

In [9]:

```
def squares(nums):
    for n in nums:
        yield n**2

for n in squares([1,2,3,4]):
    print(n)
```

```
1
4
9
16
```

In [10]:

```
tokens_lst =[x for x in tokens]
print(tokens_lst)
```

```
['Andrew', 'Ng', 'is', 'datajango', "'s", 'guru', ',', 'is', "n't", 'it',
'.', 'Anderw', 'is', 'a', 'professor', 'in', 'Stanford', 'University', ',',
'San', 'Francisco', ',', 'California', '.']
```

In [ ]:

```
len(tokens_lst)
```

In [ ]:

```
data = pd.read_csv('C:/Users/thisi/Workspace/AI_dataset/HPC_data/health_care_news_data.csv'
```

In [14]:

```
data.shape
```

Out[14]:

```
(8577, 12)
```

In [13]:

```
data.columns
```

Out[13]:

```
Index(['_id', 'crawlDate', 'headline', 'embargoDate', 'sourceUrl',
       'description', 'cureType', 'publishedDate', 'isLead', 'state', 'docI
d',
       'id'],
      dtype='object')
```

In [15]:

```
data_set = data[['headline','description']]
```

In [16]:

```
data_set.shape
```

Out[16]:

```
(8577, 2)
```

Analyse data: Let's have a look at one article

- Observation 1: We have some special characters and words such as '\n', '“', '\t', 'PF-06651600', etc.
- We should definetly remore special charactes such as '\n', '“', '\t' before we proceed further.

In [18]:

```
sample_text = data_set['description'][0]
print(sample_text)
```

```
2017-12-14 00:00:00

Global program to commence with pivotal study B7451012 in North America, Aus
tralia and Europe; broader regional rollout in 2018
Thursday, December 14, 2017 - 7:30amESTPfizer Inc. (NYSE:PFE) today announce
d the initiation of a Phase 3 program for its once-daily Janus kinase 1 (JAK
1) inhibitor PF-04965842, to evaluate the efficacy and safety of PF-04965842
for the treatment of moderate-to-severe atopic dermatitis (AD). This is the
first trial in the JAK1 Atopic Dermatitis Efficacy and safety (JADE) global
development program.
“By initiating this Phase 3 program in atopic dermatitis, we hope to provide
a new potential treatment option for people suffering with this condition,”
said Michael Corbo, Chief Development Officer, Inflammation & Immunology, Pf
izer Global Product Development. “Pfizer continues to build a leadership pos
ition in inflammation and immunology research with the advancement of this i
mportant, Pfizer-discovered investigational oral JAK1 inhibitor.”
About the Phase 3 Trial B7451012
This Phase 3 trial is a randomized, double-blind, placebo-controlled, parall
el-group study designed to evaluate the efficacy and safety of PF-04965842 i
n 375 patients 12 years and older with moderate-to-severe AD. Trial particip
ants will be randomly assigned to receive 200 mg or 100 mg once daily or pla
cebo.
The primary endpoints are the proportion of patients achieving an Investigat
or Global Assessment (IGA) score of 0/1 and ≥2 point improvement, and the pr
oportion of patients with at least a 75% or greater change from baseline in
their Eczema Area and Severity Index (EASI) score. Key secondary endpoints i
nclude the pruritus numerical rating scale, the Pruritus and Symptoms Assess
ment for Atopic Dermatitis (PSAAD) electronic diary and safety measures such
as the incidence of treatment emergent adverse events and laboratory abnorma
lities. The treatment duration will be 12 weeks, the same duration as the Ph
ase 2b study B7451006, with a 4 week safety follow-up period or the option t
o enter a long-term extension study (B7451015) at Week 12. More on the study
can be found on www.clinicaltrials.gov under the identifier NCT03349060.
The design of the Phase 3 trial is based on the Phase 2 results that were pr
esented at the 26th Congress of the European Academy of Dermatology and Vene
reology in September 2017.
About Atopic Dermatitis
Atopic dermatitis, also commonly called atopic eczema, is inflammation of th
e skin and characterized by erythema (redness), itching (pruritus), indurati
on (hardening)/papulation (formation of papules), and oozing/crusting.
About Pfizer’s Kinase Inhibitor Leadership
PF-04965842 is an oral small molecule that selectively inhibits Janus kinase
1 (JAK1). Inhibition of JAK1 modulates multiple cytokines involved in pathop
hysiology of AD including interleukin (IL)-4, IL-13, IL-31 and interferon ga
mma (IFNɣ).
“We look forward to advancing other kinase inhibitors currently in mid-stage
research for other diseases such as alopecia, psoriasis, inflammatory bowel
disease, and rheumatoid arthritis,” said Michael Vincent, M.D, Ph.D., Senior
Vice President and Chief Scientific Officer of Inflammation and Immunology R
esearch at Pfizer.
Pfizer has established a leading kinase research capability with multiple un
ique kinase inhibitor therapies in development. As a pioneer in JAK science,
the Company is advancing several investigational programs with novel selecti
vity profiles, which, if successful, could potentially deliver transformativ
e therapies for patients. Pfizer has three additional kinase inhibitors in P
hase 2 development across multiple indications:
```

PF-06651600: A JAK3 inhibitor under investigation for the tr
eatment of rheumatoid arthritis, ulcerative colitis and alopecia areata

PF-06700841: A tyrosine kinase 2 (TYK2)/JAK1 inhibitor under
investigation for the treatment of psoriasis, ulcerative colitis and alopeci
a areata

PF-06650833: An interleukin-1 receptor-associated kinase 4
(IRAK4) inhibitor under investigation for the treatment of rheumatoid arthri
tis
Working together for a healthier world®
At Pfizer, we apply science and our global resources to bring therapies to p
eople that extend and significantly improve their lives. We strive to set th
e standard for quality, safety and value in the discovery, development and m
anufacture of health care products. Our global portfolio includes medicines
and vaccines as well as many of the world's best-known consumer health care
products. Every day, Pfizer colleagues work across developed and emerging ma
rkets to advance wellness, prevention, treatments and cures that challenge t
he most feared diseases of our time. Consistent with our responsibility as o
ne of the world's premier innovative biopharmaceutical companies, we collabo
rate with health care providers, governments and local communities to suppor
t and expand access to reliable, affordable health care around the world. Fo
r more than 150 years, we have worked to make a difference for all who rely
on us. We routinely post information that may be important to investors on o
ur website at www.pfizer.com. In addition, to learn more, please visit us on
www.pfizer.com and follow us on Twitter at @Pfizer and @Pfizer_News, LinkedI
n, YouTube and like us on Facebook at Facebook.com/Pfizer.
DISCLOSURE NOTICE: The information contained in this release is as of Decemb
er 12, 2017. Pfizer assumes no obligation to update forward-looking statemen
ts contained in this release as the result of new information or future even
ts or developments.
This release contains forward-looking information about PF-04965842 and Pfiz
er’s ongoing investigational programs in kinase inhibitor therapies, includi
ng their potential benefits, that involves substantial risks and uncertainti
es that could cause actual results to differ materially from those expressed
or implied by such statements. Risks and uncertainties include, among other
things, the uncertainties inherent in research and development, including th
e ability to meet anticipated clinical trial commencement and completion dat
es and regulatory submission dates, as well as the possibility of unfavorabl
e clinical trial results, including unfavorable new clinical data and additi
onal analyses of existing data; risks associated with preliminary data; the
risk that clinical trial data are subject to differing interpretations, and,
even when we view data as sufficient to support the safety and/or effectiven
ess of a product candidate, regulatory authorities may not share our views a
nd may require additional data or may deny approval altogether; whether regu
latory authorities will be satisfied with the design of and results from our
clinical studies; whether and when drug applications may be filed in any jur
isdictions for any potential indication for PF-04965842 or any other investi
gational kinase inhibitor therapies; whether and when any such applications
may be approved by regulatory authorities, which will depend on the assessme
nt by such regulatory authorities of the benefit-risk profile suggested by t
he totality of the efficacy and safety information submitted, and, if approv
ed, whether PF-04965842 or any such other investigational kinase inhibitor t
herapies will be commercially successful; decisions by regulatory authoritie
s regarding labeling, safety and other matters that could affect the availab
ility or commercial potential of PF-04965842 or any other investigational ki
nase inhibitor therapies; and competitive developments.
A further description of risks and uncertainties can be found in Pfizer’s An
nual Report on Form 10-K for the fiscal year ended December 31, 2016 and in
its subsequent reports on Form 10-Q, including in the sections thereof capti

```
oned “Risk Factors” and “Forward-Looking Information and Factors That May Af
fect Future Results”, as well as in its subsequent reports on Form 8-K, all
of which are filed with the U.S. Securities and Exchange Commission and avai
lable at www.sec.gov and www.pfizer.com.
Contact: Media:
        Neha Wadhwa, 212-733-2835[email protected]
        or
        Investors:
        Ryan Crowe, 212-733-8160[email protected]
```

Let's find non-alpha/neumerics

In [34]:

```
search_result = re.findall('[\t\n\r\f\v]', sample_text)
```

In [35]:

```
search_result
```

Out[35]:

```
['\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\t',
 '\t',
 '\n',
 '\n',
 '\t',
 '\t',
 '\n',
 '\n',
 '\t',
 '\t',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\n',
 '\t',
 '\n',
 '\t',
 '\n',
 '\t',
 '\n',
 '\t']
```

In [36]:

```
sample_text = re.sub('[\t\n\r\f\v\"]', ' ',sample_text)
```

In [37]:

```
re.findall('[\t\n\r\f\v]', sample_text)
```

Out[37]:

```
[]
```

In [38]:

```
sample_text
```

Out[38]:

"2017-12-14 00:00:00 Global program to commence with pivotal study B7451012 in North America, Australia and Europe; broader regional rollout in 2018 Thursday, December 14, 2017 - 7:30amESTPfizer Inc. (NYSE:PFE) today announced the initiation of a Phase 3 program for its once-daily Janus kinase 1 (JAK1) inhibitor PF-04965842, to evaluate the efficacy and safety of PF-04965842 for the treatment of moderate-to-severe atopic dermatitis (AD). This is the first trial in the JAK1 Atopic Dermatitis Efficacy and safety (JADE) global development program. “By initiating this Phase 3 program in atopic dermatitis, we hope to provide a new potential treatment option for people suffering with this condition,” said Michael Corbo, Chief Development Officer, Inflammation & Immunology, Pfizer Global Product Development. “Pfizer continues to build a leadership position in inflammation and immunology research with the advancement of this important, Pfizer-discovered investigational oral JAK1 inhibitor.” About the Phase 3 Trial B7451012 This Phase 3 trial is a randomized, double-blind, placebo-controlled, parallel-group study designed to evaluate the efficacy and safety of PF-04965842 in 375 patients 12 years and older with moderate-to-severe AD. Trial participants will be randomly assigned to receive 200 mg or 100 mg once daily or placebo. The primary endpoints are the proportion of patients achieving an Investigator Global Assessment (IGA) score of 0/1 and ≥2 point improvement, and the proportion of patients with at least a 75% or greater change from baseline in their Eczema Area and Severity Index (EASI) score. Key secondary endpoints include the pruritus numerical rating scale, the Pruritus and Symptoms Assessment for Atopic Dermatitis (PSAAD) electronic diary and safety measures such as the incidence of treatment emergent adverse events and laboratory abnormalities. The treatment duration will be 12 weeks, the same duration as the Phase 2b study B7451006, with a 4 week safety follow-up period or the option to enter a long-term extension study (B7451015) at Week 12. More on the study can be found on www.clinicaltrials.gov under the identifier NCT03349060. The design of the Phase 3 trial is based on the Phase 2 results that were presented at the 26th Congress of the European Academy of Dermatology and Venereology in September 2017. About Atopic Dermatitis Atopic dermatitis, also commonly called atopic eczema, is inflammation of the skin and characterized by erythema (redness), itching (pruritus), induration (hardening)/papulation (formation of papules), and oozing/crusting. About Pfizer’s Kinase Inhibitor Leadership PF-04965842 is an oral small molecule that selectively inhibits Janus kinase 1 (JAK1). Inhibition of JAK1 modulates multiple cytokines involved in pathophysiology of AD including interleukin (IL)-4, IL-13, IL-31 and interferon gamma (IFNɣ). “We look forward to advancing other kinase inhibitors currently in mid-stage research for other diseases such as alopecia, psoriasis, inflammatory bowel disease, and rheumatoid arthritis,” said Michael Vincent, M.D, Ph.D., Senior Vice President and Chief Scientific Officer of Inflammation and Immunology Research at Pfizer. Pfizer has established a leading kinase research capability with multiple unique kinase inhibitor therapies in development. As a pioneer in JAK science, the Company is advancing several investigational programs with novel selectivity profiles, which, if successful, could potentially deliver transformative therapies for patients. Pfizer has three additional kinase inhibitors in Phase 2 development across multiple indications: PF-06651600: A JAK3 inhibitor under investigation for the treatment of rheumatoid arthritis, ulcerative colitis and alopecia areata PF-06700841: A tyrosine kinase 2 (TYK2)/JAK1 inhibitor under investigation for the treatment of psoriasis, ulcerative colitis and alopecia areata PF-06650833: An interleukin-1 receptor-associated kinase 4 (IRAK4) inhibitor under investigation for the treatment of rheumatoid arthritis Working together for a healthier world® At Pfiz

```
er, we apply science and our global resources to bring therapies to people t
hat extend and significantly improve their lives. We strive to set the stand
ard for quality, safety and value in the discovery, development and manufact
ure of health care products. Our global portfolio includes medicines and vac
cines as well as many of the world's best-known consumer health care product
s. Every day, Pfizer colleagues work across developed and emerging markets t
o advance wellness, prevention, treatments and cures that challenge the most
feared diseases of our time. Consistent with our responsibility as one of th
e world's premier innovative biopharmaceutical companies, we collaborate wit
h health care providers, governments and local communities to support and ex
pand access to reliable, affordable health care around the world. For more t
han 150 years, we have worked to make a difference for all who rely on us. W
e routinely post information that may be important to investors on our websi
te at www.pfizer.com. In addition, to learn more, please visit us on www.pfi
zer.com and follow us on Twitter at @Pfizer and @Pfizer_News, LinkedIn, YouT
ube and like us on Facebook at Facebook.com/Pfizer. DISCLOSURE NOTICE: The i
nformation contained in this release is as of December 12, 2017. Pfizer assu
mes no obligation to update forward-looking statements contained in this rel
ease as the result of new information or future events or developments. This
release contains forward-looking information about PF-04965842 and Pfizer’s
ongoing investigational programs in kinase inhibitor therapies, including th
eir potential benefits, that involves substantial risks and uncertainties th
at could cause actual results to differ materially from those expressed or i
mplied by such statements. Risks and uncertainties include, among other thin
gs, the uncertainties inherent in research and development, including the ab
ility to meet anticipated clinical trial commencement and completion dates a
nd regulatory submission dates, as well as the possibility of unfavorable cl
inical trial results, including unfavorable new clinical data and additional
analyses of existing data; risks associated with preliminary data; the risk
that clinical trial data are subject to differing interpretations, and, even
when we view data as sufficient to support the safety and/or effectiveness o
f a product candidate, regulatory authorities may not share our views and ma
y require additional data or may deny approval altogether; whether regulator
y authorities will be satisfied with the design of and results from our clin
ical studies; whether and when drug applications may be filed in any jurisdi
ctions for any potential indication for PF-04965842 or any other investigati
onal kinase inhibitor therapies; whether and when any such applications may
be approved by regulatory authorities, which will depend on the assessment b
y such regulatory authorities of the benefit-risk profile suggested by the t
otality of the efficacy and safety information submitted, and, if approved,
whether PF-04965842 or any such other investigational kinase inhibitor thera
pies will be commercially successful; decisions by regulatory authorities re
garding labeling, safety and other matters that could affect the availabilit
y or commercial potential of PF-04965842 or any other investigational kinase
inhibitor therapies; and competitive developments. A further description of
risks and uncertainties can be found in Pfizer’s Annual Report on Form 10-K
for the fiscal year ended December 31, 2016 and in its subsequent reports on
Form 10-Q, including in the sections thereof captioned “Risk Factors” and “F
orward-Looking Information and Factors That May Affect Future Results”, as w
ell as in its subsequent reports on Form 8-K, all of which are filed with th
e U.S. Securities and Exchange Commission and available at www.sec.gov and w
ww.pfizer.com. Contact: Media:  Neha Wadhwa, 212-733-2835[email protected]
or  Investors:  Ryan Crowe, 212-733-8160[email protected]"
```

In [39]:

```
re.findall(r'Contact:[\s\w,\-\d\[\]\:]*', sample_text, re.I)
```

Out[39]:

```
['Contact: Media:  Neha Wadhwa, 212-733-2835[email protected]  or  Investor
s:  Ryan Crowe, 212-733-8160[email protected]']
```

In [40]:

```
sample_text = re.sub(r'Contact:[\s\w,\-\d\[\]\:]*', ' ',sample_text)
```

In [41]:

```
sample_text
```

Out[41]:

```
"2017-12-14 00:00:00  Global program to commence with pivotal study B7451012
in North America, Australia and Europe; broader regional rollout in 2018 Thu
rsday, December 14, 2017 - 7:30amESTPfizer Inc. (NYSE:PFE) today announced t
he initiation of a Phase 3 program for its once-daily Janus kinase 1 (JAK1)
inhibitor PF-04965842, to evaluate the efficacy and safety of PF-04965842 fo
r the treatment of moderate-to-severe atopic dermatitis (AD). This is the fi
rst trial in the JAK1 Atopic Dermatitis Efficacy and safety (JADE) global de
velopment program. “By initiating this Phase 3 program in atopic dermatitis,
we hope to provide a new potential treatment option for people suffering wit
h this condition,” said Michael Corbo, Chief Development Officer, Inflammati
on & Immunology, Pfizer Global Product Development. “Pfizer continues to bui
ld a leadership position in inflammation and immunology research with the ad
vancement of this important, Pfizer-discovered investigational oral JAK1 inh
ibitor.” About the Phase 3 Trial B7451012 This Phase 3 trial is a randomize
d, double-blind, placebo-controlled, parallel-group study designed to evalua
te the efficacy and safety of PF-04965842 in 375 patients 12 years and older
with moderate-to-severe AD. Trial participants will be randomly assigned to
receive 200 mg or 100 mg once daily or placebo. The primary endpoints are th
e proportion of patients achieving an Investigator Global Assessment (IGA) s
core of 0/1 and ≥2 point improvement, and the proportion of patients with at
least a 75% or greater change from baseline in their Eczema Area and Severit
y Index (EASI) score. Key secondary endpoints include the pruritus numerical
rating scale, the Pruritus and Symptoms Assessment for Atopic Dermatitis (PS
AAD) electronic diary and safety measures such as the incidence of treatment
emergent adverse events and laboratory abnormalities. The treatment duration
will be 12 weeks, the same duration as the Phase 2b study B7451006, with a 4
week safety follow-up period or the option to enter a long-term extension st
udy (B7451015) at Week 12. More on the study can be found on www.clinicaltri
als.gov under the identifier NCT03349060. The design of the Phase 3 trial is
based on the Phase 2 results that were presented at the 26th Congress of the
European Academy of Dermatology and Venereology in September 2017. About Ato
pic Dermatitis Atopic dermatitis, also commonly called atopic eczema, is inf
lammation of the skin and characterized by erythema (redness), itching (prur
itus), induration (hardening)/papulation (formation of papules), and oozing/
crusting. About Pfizer’s Kinase Inhibitor Leadership PF-04965842 is an oral
small molecule that selectively inhibits Janus kinase 1 (JAK1). Inhibition o
f JAK1 modulates multiple cytokines involved in pathophysiology of AD includ
ing interleukin (IL)-4, IL-13, IL-31 and interferon gamma (IFNɣ). “We look f
orward to advancing other kinase inhibitors currently in mid-stage research
for other diseases such as alopecia, psoriasis, inflammatory bowel disease,
and rheumatoid arthritis,” said Michael Vincent, M.D, Ph.D., Senior Vice Pre
sident and Chief Scientific Officer of Inflammation and Immunology Research
at Pfizer. Pfizer has established a leading kinase research capability with
multiple unique kinase inhibitor therapies in development. As a pioneer in J
AK science, the Company is advancing several investigational programs with n
ovel selectivity profiles, which, if successful, could potentially deliver t
ransformative therapies for patients. Pfizer has three additional kinase inh
ibitors in Phase 2 development across multiple indications:    PF-06651600:
A JAK3 inhibitor under investigation for the treatment of rheumatoid arthrit
is, ulcerative colitis and alopecia areata    PF-06700841: A tyrosine kinase
2 (TYK2)/JAK1 inhibitor under investigation for the treatment of psoriasis,
ulcerative colitis and alopecia areata    PF-06650833: An interleukin-1 rece
ptor-associated kinase 4 (IRAK4) inhibitor under investigation for the treat
ment of rheumatoid arthritis Working together for a healthier world® At Pfiz
er, we apply science and our global resources to bring therapies to people t
```

```
hat extend and significantly improve their lives. We strive to set the stand
ard for quality, safety and value in the discovery, development and manufact
ure of health care products. Our global portfolio includes medicines and vac
cines as well as many of the world's best-known consumer health care product
s. Every day, Pfizer colleagues work across developed and emerging markets t
o advance wellness, prevention, treatments and cures that challenge the most
feared diseases of our time. Consistent with our responsibility as one of th
e world's premier innovative biopharmaceutical companies, we collaborate wit
h health care providers, governments and local communities to support and ex
pand access to reliable, affordable health care around the world. For more t
han 150 years, we have worked to make a difference for all who rely on us. W
e routinely post information that may be important to investors on our websi
te at www.pfizer.com. In addition, to learn more, please visit us on www.pfi
zer.com and follow us on Twitter at @Pfizer and @Pfizer_News, LinkedIn, YouT
ube and like us on Facebook at Facebook.com/Pfizer. DISCLOSURE NOTICE: The i
nformation contained in this release is as of December 12, 2017. Pfizer assu
mes no obligation to update forward-looking statements contained in this rel
ease as the result of new information or future events or developments. This
release contains forward-looking information about PF-04965842 and Pfizer’s
ongoing investigational programs in kinase inhibitor therapies, including th
eir potential benefits, that involves substantial risks and uncertainties th
at could cause actual results to differ materially from those expressed or i
mplied by such statements. Risks and uncertainties include, among other thin
gs, the uncertainties inherent in research and development, including the ab
ility to meet anticipated clinical trial commencement and completion dates a
nd regulatory submission dates, as well as the possibility of unfavorable cl
inical trial results, including unfavorable new clinical data and additional
analyses of existing data; risks associated with preliminary data; the risk
that clinical trial data are subject to differing interpretations, and, even
when we view data as sufficient to support the safety and/or effectiveness o
f a product candidate, regulatory authorities may not share our views and ma
y require additional data or may deny approval altogether; whether regulator
y authorities will be satisfied with the design of and results from our clin
ical studies; whether and when drug applications may be filed in any jurisdi
ctions for any potential indication for PF-04965842 or any other investigati
onal kinase inhibitor therapies; whether and when any such applications may
be approved by regulatory authorities, which will depend on the assessment b
y such regulatory authorities of the benefit-risk profile suggested by the t
otality of the efficacy and safety information submitted, and, if approved,
whether PF-04965842 or any such other investigational kinase inhibitor thera
pies will be commercially successful; decisions by regulatory authorities re
garding labeling, safety and other matters that could affect the availabilit
y or commercial potential of PF-04965842 or any other investigational kinase
inhibitor therapies; and competitive developments. A further description of
risks and uncertainties can be found in Pfizer’s Annual Report on Form 10-K
for the fiscal year ended December 31, 2016 and in its subsequent reports on
Form 10-Q, including in the sections thereof captioned “Risk Factors” and “F
orward-Looking Information and Factors That May Affect Future Results”, as w
ell as in its subsequent reports on Form 8-K, all of which are filed with th
e U.S. Securities and Exchange Commission and available at www.sec.gov and w
ww.pfizer.com.  "
```

Tockenization & Understanding Features

Let's understand Tokenizers a little bit before applying them on our data

In [42]:

```
import nltk
from nltk.tokenize import WhitespaceTokenizer
from nltk.tokenize import RegexpTokenizer
```

In [47]:

```
#regx_tokenizer = RegexpTokenizer('\w+|\$[\d\.]+|\S+')
```

In [48]:

```
regx_tokenizer = RegexpTokenizer('\S+')
```

In [49]:

```
regx_tokens = regx_tokenizer.tokenize(sample_text)
```

In [50]:

```
print(regx_tokens)
```

```
['2017-12-14', '00:00:00', 'Global', 'program', 'to', 'commence', 'with', 'p
ivotal', 'study', 'B7451012', 'in', 'North', 'America,', 'Australia', 'and',
'Europe;', 'broader', 'regional', 'rollout', 'in', '2018', 'Thursday,', 'Dec
ember', '14,', '2017', '-', '7:30amESTPfizer', 'Inc.', '(NYSE:PFE)', 'toda
y', 'announced', 'the', 'initiation', 'of', 'a', 'Phase', '3', 'program', 'f
or', 'its', 'once-daily', 'Janus', 'kinase', '1', '(JAK1)', 'inhibitor', 'PF
-04965842,', 'to', 'evaluate', 'the', 'efficacy', 'and', 'safety', 'of', 'PF
-04965842', 'for', 'the', 'treatment', 'of', 'moderate-to-severe', 'atopic',
'dermatitis', '(AD).', 'This', 'is', 'the', 'first', 'trial', 'in', 'the',
'JAK1', 'Atopic', 'Dermatitis', 'Efficacy', 'and', 'safety', '(JADE)', 'glob
al', 'development', 'program.', '“By', 'initiating', 'this', 'Phase', '3',
'program', 'in', 'atopic', 'dermatitis,', 'we', 'hope', 'to', 'provide',
'a', 'new', 'potential', 'treatment', 'option', 'for', 'people', 'sufferin
g', 'with', 'this', 'condition,”', 'said', 'Michael', 'Corbo,', 'Chief', 'De
velopment', 'Officer,', 'Inflammation', '&', 'Immunology,', 'Pfizer', 'Globa
l', 'Product', 'Development.', '“Pfizer', 'continues', 'to', 'build', 'a',
'leadership', 'position', 'in', 'inflammation', 'and', 'immunology', 'resear
ch', 'with', 'the', 'advancement', 'of', 'this', 'important,', 'Pfizer-disco
vered', 'investigational', 'oral', 'JAK1', 'inhibitor.”', 'About', 'the', 'P
hase', '3', 'Trial', 'B7451012', 'This', 'Phase', '3', 'trial', 'is', 'a',
'randomized,', 'double-blind,', 'placebo-controlled,', 'parallel-group', 'st
udy', 'designed', 'to', 'evaluate', 'the', 'efficacy', 'and', 'safety', 'o
f', 'PF-04965842', 'in', '375', 'patients', '12', 'years', 'and', 'older',
'with', 'moderate-to-severe', 'AD.', 'Trial', 'participants', 'will', 'be',
'randomly', 'assigned', 'to', 'receive', '200', 'mg', 'or', '100', 'mg', 'on
ce', 'daily', 'or', 'placebo.', 'The', 'primary', 'endpoints', 'are', 'the',
'proportion', 'of', 'patients', 'achieving', 'an', 'Investigator', 'Global',
'Assessment', '(IGA)', 'score', 'of', '0/1', 'and', '≥2', 'point', 'improvem
ent,', 'and', 'the', 'proportion', 'of', 'patients', 'with', 'at', 'least',
'a', '75%', 'or', 'greater', 'change', 'from', 'baseline', 'in', 'their', 'E
czema', 'Area', 'and', 'Severity', 'Index', '(EASI)', 'score.', 'Key', 'seco
ndary', 'endpoints', 'include', 'the', 'pruritus', 'numerical', 'rating', 's
cale,', 'the', 'Pruritus', 'and', 'Symptoms', 'Assessment', 'for', 'Atopic',
'Dermatitis', '(PSAAD)', 'electronic', 'diary', 'and', 'safety', 'measures',
'such', 'as', 'the', 'incidence', 'of', 'treatment', 'emergent', 'adverse',
'events', 'and', 'laboratory', 'abnormalities.', 'The', 'treatment', 'durati
on', 'will', 'be', '12', 'weeks,', 'the', 'same', 'duration', 'as', 'the',
'Phase', '2b', 'study', 'B7451006,', 'with', 'a', '4', 'week', 'safety', 'fo
llow-up', 'period', 'or', 'the', 'option', 'to', 'enter', 'a', 'long-term',
'extension', 'study', '(B7451015)', 'at', 'Week', '12.', 'More', 'on', 'th
e', 'study', 'can', 'be', 'found', 'on', 'www.clinicaltrials.gov', 'under',
'the', 'identifier', 'NCT03349060.', 'The', 'design', 'of', 'the', 'Phase',
'3', 'trial', 'is', 'based', 'on', 'the', 'Phase', '2', 'results', 'that',
'were', 'presented', 'at', 'the', '26th', 'Congress', 'of', 'the', 'Europea
n', 'Academy', 'of', 'Dermatology', 'and', 'Venereology', 'in', 'September',
'2017.', 'About', 'Atopic', 'Dermatitis', 'Atopic', 'dermatitis,', 'also',
'commonly', 'called', 'atopic', 'eczema,', 'is', 'inflammation', 'of', 'th
e', 'skin', 'and', 'characterized', 'by', 'erythema', '(redness),', 'itchin
g', '(pruritus),', 'induration', '(hardening)/papulation', '(formation', 'o
f', 'papules),', 'and', 'oozing/crusting.', 'About', 'Pfizer’s', 'Kinase',
'Inhibitor', 'Leadership', 'PF-04965842', 'is', 'an', 'oral', 'small', 'mole
cule', 'that', 'selectively', 'inhibits', 'Janus', 'kinase', '1', '(JAK1).',
'Inhibition', 'of', 'JAK1', 'modulates', 'multiple', 'cytokines', 'involve
d', 'in', 'pathophysiology', 'of', 'AD', 'including', 'interleukin', '(IL)-
4,', 'IL-13,', 'IL-31', 'and', 'interferon', 'gamma', '(IFNɣ).', '“We', 'loo
k', 'forward', 'to', 'advancing', 'other', 'kinase', 'inhibitors', 'currentl
y', 'in', 'mid-stage', 'research', 'for', 'other', 'diseases', 'such', 'as',
```

'alopecia,', 'psoriasis,', 'inflammatory', 'bowel', 'disease,', 'and', 'rheu
matoid', 'arthritis,”', 'said', 'Michael', 'Vincent,', 'M.D,', 'Ph.D.,', 'Se
nior', 'Vice', 'President', 'and', 'Chief', 'Scientific', 'Officer', 'of',
'Inflammation', 'and', 'Immunology', 'Research', 'at', 'Pfizer.', 'Pfizer',
'has', 'established', 'a', 'leading', 'kinase', 'research', 'capability', 'w
ith', 'multiple', 'unique', 'kinase', 'inhibitor', 'therapies', 'in', 'devel
opment.', 'As', 'a', 'pioneer', 'in', 'JAK', 'science,', 'the', 'Company',
'is', 'advancing', 'several', 'investigational', 'programs', 'with', 'nove
l', 'selectivity', 'profiles,', 'which,', 'if', 'successful,', 'could', 'pot
entially', 'deliver', 'transformative', 'therapies', 'for', 'patients.', 'Pf
izer', 'has', 'three', 'additional', 'kinase', 'inhibitors', 'in', 'Phase',
'2', 'development', 'across', 'multiple', 'indications:', 'PF-06651600:',
'A', 'JAK3', 'inhibitor', 'under', 'investigation', 'for', 'the', 'treatmen
t', 'of', 'rheumatoid', 'arthritis,', 'ulcerative', 'colitis', 'and', 'alope
cia', 'areata', 'PF-06700841:', 'A', 'tyrosine', 'kinase', '2', '(TYK2)/JAK
1', 'inhibitor', 'under', 'investigation', 'for', 'the', 'treatment', 'of',
'psoriasis,', 'ulcerative', 'colitis', 'and', 'alopecia', 'areata', 'PF-0665
0833:', 'An', 'interleukin-1', 'receptor-associated', 'kinase', '4', '(IRAK
4)', 'inhibitor', 'under', 'investigation', 'for', 'the', 'treatment', 'of',
'rheumatoid', 'arthritis', 'Working', 'together', 'for', 'a', 'healthier',
'world®', 'At', 'Pfizer,', 'we', 'apply', 'science', 'and', 'our', 'global',
'resources', 'to', 'bring', 'therapies', 'to', 'people', 'that', 'extend',
'and', 'significantly', 'improve', 'their', 'lives.', 'We', 'strive', 'to',
'set', 'the', 'standard', 'for', 'quality,', 'safety', 'and', 'value', 'in',
'the', 'discovery,', 'development', 'and', 'manufacture', 'of', 'health', 'c
are', 'products.', 'Our', 'global', 'portfolio', 'includes', 'medicines', 'a
nd', 'vaccines', 'as', 'well', 'as', 'many', 'of', 'the', "world's", 'best-k
nown', 'consumer', 'health', 'care', 'products.', 'Every', 'day,', 'Pfizer',
'colleagues', 'work', 'across', 'developed', 'and', 'emerging', 'markets',
'to', 'advance', 'wellness,', 'prevention,', 'treatments', 'and', 'cures',
'that', 'challenge', 'the', 'most', 'feared', 'diseases', 'of', 'our', 'tim
e.', 'Consistent', 'with', 'our', 'responsibility', 'as', 'one', 'of', 'th
e', "world's", 'premier', 'innovative', 'biopharmaceutical', 'companies,',
'we', 'collaborate', 'with', 'health', 'care', 'providers,', 'governments',
'and', 'local', 'communities', 'to', 'support', 'and', 'expand', 'access',
'to', 'reliable,', 'affordable', 'health', 'care', 'around', 'the', 'worl
d.', 'For', 'more', 'than', '150', 'years,', 'we', 'have', 'worked', 'to',
'make', 'a', 'difference', 'for', 'all', 'who', 'rely', 'on', 'us.', 'We',
'routinely', 'post', 'information', 'that', 'may', 'be', 'important', 'to',
'investors', 'on', 'our', 'website', 'at', 'www.pfizer.com.', 'In', 'additio
n,', 'to', 'learn', 'more,', 'please', 'visit', 'us', 'on', 'www.pfizer.co
m', 'and', 'follow', 'us', 'on', 'Twitter', 'at', '@Pfizer', 'and', '@Pfizer
_News,', 'LinkedIn,', 'YouTube', 'and', 'like', 'us', 'on', 'Facebook', 'a
t', 'Facebook.com/Pfizer.', 'DISCLOSURE', 'NOTICE:', 'The', 'information',
'contained', 'in', 'this', 'release', 'is', 'as', 'of', 'December', '12,',
'2017.', 'Pfizer', 'assumes', 'no', 'obligation', 'to', 'update', 'forward-l
ooking', 'statements', 'contained', 'in', 'this', 'release', 'as', 'the', 'r
esult', 'of', 'new', 'information', 'or', 'future', 'events', 'or', 'develop
ments.', 'This', 'release', 'contains', 'forward-looking', 'information', 'a
bout', 'PF-04965842', 'and', 'Pfizer’s', 'ongoing', 'investigational', 'prog
rams', 'in', 'kinase', 'inhibitor', 'therapies,', 'including', 'their', 'pot
ential', 'benefits,', 'that', 'involves', 'substantial', 'risks', 'and', 'un
certainties', 'that', 'could', 'cause', 'actual', 'results', 'to', 'differ',
'materially', 'from', 'those', 'expressed', 'or', 'implied', 'by', 'such',
'statements.', 'Risks', 'and', 'uncertainties', 'include,', 'among', 'othe
r', 'things,', 'the', 'uncertainties', 'inherent', 'in', 'research', 'and',
'development,', 'including', 'the', 'ability', 'to', 'meet', 'anticipated',
'clinical', 'trial', 'commencement', 'and', 'completion', 'dates', 'and', 'r
egulatory', 'submission', 'dates,', 'as', 'well', 'as', 'the', 'possibilit
y', 'of', 'unfavorable', 'clinical', 'trial', 'results,', 'including', 'unfa
vorable', 'new', 'clinical', 'data', 'and', 'additional', 'analyses', 'of',

'existing', 'data;', 'risks', 'associated', 'with', 'preliminary', 'data;', 'the', 'risk', 'that', 'clinical', 'trial', 'data', 'are', 'subject', 'to', 'differing', 'interpretations,', 'and,', 'even', 'when', 'we', 'view', 'data', 'as', 'sufficient', 'to', 'support', 'the', 'safety', 'and/or', 'effectiveness', 'of', 'a', 'product', 'candidate,', 'regulatory', 'authorities', 'may', 'not', 'share', 'our', 'views', 'and', 'may', 'require', 'additional', 'data', 'or', 'may', 'deny', 'approval', 'altogether;', 'whether', 'regulatory', 'authorities', 'will', 'be', 'satisfied', 'with', 'the', 'design', 'of', 'and', 'results', 'from', 'our', 'clinical', 'studies;', 'whether', 'and', 'when', 'drug', 'applications', 'may', 'be', 'filed', 'in', 'any', 'jurisdictions', 'for', 'any', 'potential', 'indication', 'for', 'PF-04965842', 'or', 'any', 'other', 'investigational', 'kinase', 'inhibitor', 'therapies;', 'whether', 'and', 'when', 'any', 'such', 'applications', 'may', 'be', 'approved', 'by', 'regulatory', 'authorities,', 'which', 'will', 'depend', 'on', 'the', 'assessment', 'by', 'such', 'regulatory', 'authorities', 'of', 'the', 'benefit-risk', 'profile', 'suggested', 'by', 'the', 'totality', 'of', 'the', 'efficacy', 'and', 'safety', 'information', 'submitted,', 'and,', 'if', 'approved,', 'whether', 'PF-04965842', 'or', 'any', 'such', 'other', 'investigational', 'kinase', 'inhibitor', 'therapies', 'will', 'be', 'commercially', 'successful;', 'decisions', 'by', 'regulatory', 'authorities', 'regarding', 'labeling,', 'safety', 'and', 'other', 'matters', 'that', 'could', 'affect', 'the', 'availability', 'or', 'commercial', 'potential', 'of', 'PF-04965842', 'or', 'any', 'other', 'investigational', 'kinase', 'inhibitor', 'therapies;', 'and', 'competitive', 'developments.', 'A', 'further', 'description', 'of', 'risks', 'and', 'uncertainties', 'can', 'be', 'found', 'in', 'Pfizer’s', 'Annual', 'Report', 'on', 'Form', '10-K', 'for', 'the', 'fiscal', 'year', 'ended', 'December', '31,', '2016', 'and', 'in', 'its', 'subsequent', 'reports', 'on', 'Form', '10-Q,', 'including', 'in', 'the', 'sections', 'thereof', 'captioned', '‘“Risk', 'Factors”', 'and', '‘“Forward-Looking', 'Information', 'and', 'Factors', 'That', 'May', 'Affect', 'Future', 'Results”,', 'as', 'well', 'as', 'in', 'its', 'subsequent', 'reports', 'on', 'Form', '8-K,', 'all', 'of', 'which', 'are', 'filed', 'with', 'the', 'U.S.', 'Securities', 'and', 'Exchange', 'Commission', 'and', 'available', 'at', 'www.sec.gov', 'and', 'www.pfizer.com.']

In [51]:

```
from nltk.tokenize import SpaceTokenizer
space_tokenizer = SpaceTokenizer()
space_tokens = space_tokenizer.tokenize(sample_text)
print(space_tokens)
```

```
['2017-12-14', '00:00:00', '', 'Global', 'program', 'to', 'commence', 'wit
h', 'pivotal', 'study', 'B7451012', 'in', 'North', 'America,', 'Australia',
'and', 'Europe;', 'broader', 'regional', 'rollout', 'in', '2018', 'Thursda
y,', 'December', '14,', '2017', '-', '7:30amESTPfizer', 'Inc.', '(NYSE:PF
E)', 'today', 'announced', 'the', 'initiation', 'of', 'a', 'Phase', '3', 'pr
ogram', 'for', 'its', 'once-daily', 'Janus', 'kinase', '1', '(JAK1)', 'inhib
itor', 'PF-04965842,', 'to', 'evaluate', 'the', 'efficacy', 'and', 'safety',
'of', 'PF-04965842', 'for', 'the', 'treatment', 'of', 'moderate-to-severe',
'atopic', 'dermatitis', '(AD).', 'This', 'is', 'the', 'first', 'trial', 'i
n', 'the', 'JAK1', 'Atopic', 'Dermatitis', 'Efficacy', 'and', 'safety', '(JA
DE)', 'global', 'development', 'program.', '"By', 'initiating', 'this', 'Pha
se', '3', 'program', 'in', 'atopic', 'dermatitis,', 'we', 'hope', 'to', 'pro
vide', 'a', 'new', 'potential', 'treatment', 'option', 'for', 'people', 'suf
fering', 'with', 'this', 'condition,"', 'said', 'Michael', 'Corbo,', 'Chie
f', 'Development', 'Officer,', 'Inflammation', '&', 'Immunology,', 'Pfizer',
'Global', 'Product', 'Development.', '"Pfizer', 'continues', 'to', 'build',
'a', 'leadership', 'position', 'in', 'inflammation', 'and', 'immunology', 'r
esearch', 'with', 'the', 'advancement', 'of', 'this', 'important,', 'Pfizer-
discovered', 'investigational', 'oral', 'JAK1', 'inhibitor."', 'About', 'th
e', 'Phase', '3', 'Trial', 'B7451012', 'This', 'Phase', '3', 'trial', 'is',
'a', 'randomized,', 'double-blind,', 'placebo-controlled,', 'parallel-grou
p', 'study', 'designed', 'to', 'evaluate', 'the', 'efficacy', 'and', 'safet
y', 'of', 'PF-04965842', 'in', '375', 'patients', '12', 'years', 'and', 'old
er', 'with', 'moderate-to-severe', 'AD.', 'Trial', 'participants', 'will',
'be', 'randomly', 'assigned', 'to', 'receive', '200', 'mg', 'or', '100', 'm
g', 'once', 'daily', 'or', 'placebo.', 'The', 'primary', 'endpoints', 'are',
'the', 'proportion', 'of', 'patients', 'achieving', 'an', 'Investigator', 'G
lobal', 'Assessment', '(IGA)', 'score', 'of', '0/1', 'and', '≥2', 'point',
'improvement,', 'and', 'the', 'proportion', 'of', 'patients', 'with', 'at',
'least', 'a', '75%', 'or', 'greater', 'change', 'from', 'baseline', 'in', 't
heir', 'Eczema', 'Area', 'and', 'Severity', 'Index', '(EASI)', 'score.', 'Ke
y', 'secondary', 'endpoints', 'include', 'the', 'pruritus', 'numerical', 'ra
ting', 'scale,', 'the', 'Pruritus', 'and', 'Symptoms', 'Assessment', 'for',
'Atopic', 'Dermatitis', '(PSAAD)', 'electronic', 'diary', 'and', 'safety',
'measures', 'such', 'as', 'the', 'incidence', 'of', 'treatment', 'emergent',
'adverse', 'events', 'and', 'laboratory', 'abnormalities.', 'The', 'treatmen
t', 'duration', 'will', 'be', '12', 'weeks,', 'the', 'same', 'duration', 'a
s', 'the', 'Phase', '2b', 'study', 'B7451006,', 'with', 'a', '4', 'week', 's
afety', 'follow-up', 'period', 'or', 'the', 'option', 'to', 'enter', 'a', 'l
ong-term', 'extension', 'study', '(B7451015)', 'at', 'Week', '12.', 'More',
'on', 'the', 'study', 'can', 'be', 'found', 'on', 'www.clinicaltrials.gov',
'under', 'the', 'identifier', 'NCT03349060.', 'The', 'design', 'of', 'the',
'Phase', '3', 'trial', 'is', 'based', 'on', 'the', 'Phase', '2', 'results',
'that', 'were', 'presented', 'at', 'the', '26th', 'Congress', 'of', 'the',
'European', 'Academy', 'of', 'Dermatology', 'and', 'Venereology', 'in', 'Sep
tember', '2017.', 'About', 'Atopic', 'Dermatitis', 'Atopic', 'dermatitis,',
'also', 'commonly', 'called', 'atopic', 'eczema,', 'is', 'inflammation', 'o
f', 'the', 'skin', 'and', 'characterized', 'by', 'erythema', '(redness),',
'itching', '(pruritus),', 'induration', '(hardening)/papulation', '(formatio
n', 'of', 'papules),', 'and', 'oozing/crusting.', 'About', 'Pfizer's', 'Kina
se', 'Inhibitor', 'Leadership', 'PF-04965842', 'is', 'an', 'oral', 'small',
'molecule', 'that', 'selectively', 'inhibits', 'Janus', 'kinase', '1', '(JAK
1).', 'Inhibition', 'of', 'JAK1', 'modulates', 'multiple', 'cytokines', 'inv
olved', 'in', 'pathophysiology', 'of', 'AD', 'including', 'interleukin', '(I
```

```
L)-4,', 'IL-13,', 'IL-31', 'and', 'interferon', 'gamma', '(IFNɣ).', '“We',
'look', 'forward', 'to', 'advancing', 'other', 'kinase', 'inhibitors', 'curr
ently', 'in', 'mid-stage', 'research', 'for', 'other', 'diseases', 'such',
'as', 'alopecia,', 'psoriasis,', 'inflammatory', 'bowel', 'disease,', 'and',
'rheumatoid', 'arthritis,”', 'said', 'Michael', 'Vincent,', 'M.D,', 'Ph.
D.,', 'Senior', 'Vice', 'President', 'and', 'Chief', 'Scientific', 'Office
r', 'of', 'Inflammation', 'and', 'Immunology', 'Research', 'at', 'Pfizer.',
'Pfizer', 'has', 'established', 'a', 'leading', 'kinase', 'research', 'capab
ility', 'with', 'multiple', 'unique', 'kinase', 'inhibitor', 'therapies', 'i
n', 'development.', 'As', 'a', 'pioneer', 'in', 'JAK', 'science,', 'the', 'C
ompany', 'is', 'advancing', 'several', 'investigational', 'programs', 'wit
h', 'novel', 'selectivity', 'profiles,', 'which,', 'if', 'successful,', 'cou
ld', 'potentially', 'deliver', 'transformative', 'therapies', 'for', 'patien
ts.', 'Pfizer', 'has', 'three', 'additional', 'kinase', 'inhibitors', 'in',
'Phase', '2', 'development', 'across', 'multiple', 'indications:', '', '',
'', 'PF-06651600:', 'A', 'JAK3', 'inhibitor', 'under', 'investigation', 'fo
r', 'the', 'treatment', 'of', 'rheumatoid', 'arthritis,', 'ulcerative', 'col
itis', 'and', 'alopecia', 'areata', '', '', '', 'PF-06700841:', 'A', 'tyrosi
ne', 'kinase', '2', '(TYK2)/JAK1', 'inhibitor', 'under', 'investigation', 'f
or', 'the', 'treatment', 'of', 'psoriasis,', 'ulcerative', 'colitis', 'and',
'alopecia', 'areata', '', '', '', 'PF-06650833:', 'An', 'interleukin-1', 're
ceptor-associated', 'kinase', '4', '(IRAK4)', 'inhibitor', 'under', 'investi
gation', 'for', 'the', 'treatment', 'of', 'rheumatoid', 'arthritis', 'Workin
g', 'together', 'for', 'a', 'healthier', 'world®', 'At', 'Pfizer,', 'we', 'a
pply', 'science', 'and', 'our', 'global', 'resources', 'to', 'bring', 'thera
pies', 'to', 'people', 'that', 'extend', 'and', 'significantly', 'improve',
'their', 'lives.', 'We', 'strive', 'to', 'set', 'the', 'standard', 'for', 'q
uality,', 'safety', 'and', 'value', 'in', 'the', 'discovery,', 'developmen
t', 'and', 'manufacture', 'of', 'health', 'care', 'products.', 'Our', 'globa
l', 'portfolio', 'includes', 'medicines', 'and', 'vaccines', 'as', 'well',
'as', 'many', 'of', 'the', "world's", 'best-known', 'consumer', 'health', 'c
are', 'products.', 'Every', 'day,', 'Pfizer', 'colleagues', 'work', 'acros
s', 'developed', 'and', 'emerging', 'markets', 'to', 'advance', 'wellness,',
'prevention,', 'treatments', 'and', 'cures', 'that', 'challenge', 'the', 'mo
st', 'feared', 'diseases', 'of', 'our', 'time.', 'Consistent', 'with', 'ou
r', 'responsibility', 'as', 'one', 'of', 'the', "world's", 'premier', 'innov
ative', 'biopharmaceutical', 'companies,', 'we', 'collaborate', 'with', 'hea
lth', 'care', 'providers,', 'governments', 'and', 'local', 'communities', 't
o', 'support', 'and', 'expand', 'access', 'to', 'reliable,', 'affordable',
'health', 'care', 'around', 'the', 'world.', 'For', 'more', 'than', '150',
'years,', 'we', 'have', 'worked', 'to', 'make', 'a', 'difference', 'for', 'a
ll', 'who', 'rely', 'on', 'us.', 'We', 'routinely', 'post', 'information',
'that', 'may', 'be', 'important', 'to', 'investors', 'on', 'our', 'website',
'at', 'www.pfizer.com.', 'In', 'addition,', 'to', 'learn', 'more,', 'pleas
e', 'visit', 'us', 'on', 'www.pfizer.com', 'and', 'follow', 'us', 'on', 'Twi
tter', 'at', '@Pfizer', 'and', '@Pfizer_News,', 'LinkedIn,', 'YouTube', 'an
d', 'like', 'us', 'on', 'Facebook', 'at', 'Facebook.com/Pfizer.', 'DISCLOSUR
E', 'NOTICE:', 'The', 'information', 'contained', 'in', 'this', 'release',
'is', 'as', 'of', 'December', '12,', '2017.', 'Pfizer', 'assumes', 'no', 'ob
ligation', 'to', 'update', 'forward-looking', 'statements', 'contained', 'i
n', 'this', 'release', 'as', 'the', 'result', 'of', 'new', 'information', 'o
r', 'future', 'events', 'or', 'developments.', 'This', 'release', 'contain
s', 'forward-looking', 'information', 'about', 'PF-04965842', 'and', 'Pfize
r’s', 'ongoing', 'investigational', 'programs', 'in', 'kinase', 'inhibitor',
'therapies,', 'including', 'their', 'potential', 'benefits,', 'that', 'invol
ves', 'substantial', 'risks', 'and', 'uncertainties', 'that', 'could', 'caus
e', 'actual', 'results', 'to', 'differ', 'materially', 'from', 'those', 'exp
ressed', 'or', 'implied', 'by', 'such', 'statements.', 'Risks', 'and', 'unce
rtainties', 'include,', 'among', 'other', 'things,', 'the', 'uncertainties',
'inherent', 'in', 'research', 'and', 'development,', 'including', 'the', 'ab
ility', 'to', 'meet', 'anticipated', 'clinical', 'trial', 'commencement', 'a
```

```
nd', 'completion', 'dates', 'and', 'regulatory', 'submission', 'dates,', 'a
s', 'well', 'as', 'the', 'possibility', 'of', 'unfavorable', 'clinical', 'tr
ial', 'results,', 'including', 'unfavorable', 'new', 'clinical', 'data', 'an
d', 'additional', 'analyses', 'of', 'existing', 'data;', 'risks', 'associate
d', 'with', 'preliminary', 'data;', 'the', 'risk', 'that', 'clinical', 'tria
l', 'data', 'are', 'subject', 'to', 'differing', 'interpretations,', 'and,',
'even', 'when', 'we', 'view', 'data', 'as', 'sufficient', 'to', 'support',
'the', 'safety', 'and/or', 'effectiveness', 'of', 'a', 'product', 'candidat
e,', 'regulatory', 'authorities', 'may', 'not', 'share', 'our', 'views', 'an
d', 'may', 'require', 'additional', 'data', 'or', 'may', 'deny', 'approval',
'altogether;', 'whether', 'regulatory', 'authorities', 'will', 'be', 'satisf
ied', 'with', 'the', 'design', 'of', 'and', 'results', 'from', 'our', 'clini
cal', 'studies;', 'whether', 'and', 'when', 'drug', 'applications', 'may',
'be', 'filed', 'in', 'any', 'jurisdictions', 'for', 'any', 'potential', 'ind
ication', 'for', 'PF-04965842', 'or', 'any', 'other', 'investigational', 'ki
nase', 'inhibitor', 'therapies;', 'whether', 'and', 'when', 'any', 'such',
'applications', 'may', 'be', 'approved', 'by', 'regulatory', 'authorities,',
'which', 'will', 'depend', 'on', 'the', 'assessment', 'by', 'such', 'regulat
ory', 'authorities', 'of', 'the', 'benefit-risk', 'profile', 'suggested', 'b
y', 'the', 'totality', 'of', 'the', 'efficacy', 'and', 'safety', 'informatio
n', 'submitted,', 'and,', 'if', 'approved,', 'whether', 'PF-04965842', 'or',
'any', 'such', 'other', 'investigational', 'kinase', 'inhibitor', 'therapie
s', 'will', 'be', 'commercially', 'successful;', 'decisions', 'by', 'regulat
ory', 'authorities', 'regarding', 'labeling,', 'safety', 'and', 'other', 'ma
tters', 'that', 'could', 'affect', 'the', 'availability', 'or', 'commercia
l', 'potential', 'of', 'PF-04965842', 'or', 'any', 'other', 'investigationa
l', 'kinase', 'inhibitor', 'therapies;', 'and', 'competitive', 'development
s.', 'A', 'further', 'description', 'of', 'risks', 'and', 'uncertainties',
'can', 'be', 'found', 'in', 'Pfizer’s', 'Annual', 'Report', 'on', 'Form', '1
0-K', 'for', 'the', 'fiscal', 'year', 'ended', 'December', '31,', '2016', 'a
nd', 'in', 'its', 'subsequent', 'reports', 'on', 'Form', '10-Q,', 'includin
g', 'in', 'the', 'sections', 'thereof', 'captioned', '“Risk', 'Factors”', 'a
nd', '“Forward-Looking', 'Information', 'and', 'Factors', 'That', 'May', 'Af
fect', 'Future', 'Results”,', 'as', 'well', 'as', 'in', 'its', 'subsequent',
'reports', 'on', 'Form', '8-K,', 'all', 'of', 'which', 'are', 'filed', 'wit
h', 'the', 'U.S.', 'Securities', 'and', 'Exchange', 'Commission', 'and', 'av
ailable', 'at', 'www.sec.gov', 'and', 'www.pfizer.com.', '', '']
```

In [52]:

```
from nltk.parse.corenlp import CoreNLPParser
```

In [53]:

```
st_parser = CoreNLPParser(url='http://localhost:9000')
st_tokens = st_parser.tokenize(sample_text)
print(st_tokens)
```

```
<generator object GenericCoreNLPParser.tokenize at 0x0000022CB2B3FE58>
```

In [54]:

```
print(type(st_tokens))
```

```
<class 'generator'>
```

In [55]:

```
st_tokens_lst =[x for x in st_tokens]
print(st_tokens_lst)
```

```
['2017-12-14', '00:00:00', 'Global', 'program', 'to', 'commence', 'with', 'p
ivotal', 'study', 'B7451012', 'in', 'North', 'America', ',', 'Australia', 'a
nd', 'Europe', ';', 'broader', 'regional', 'rollout', 'in', '2018', 'Thursda
y', ',', 'December', '14', ',', '2017', '-', '7:30', 'amESTPfizer', 'Inc.',
'(', 'NYSE', ':', 'PFE', ')', 'today', 'announced', 'the', 'initiation', 'o
f', 'a', 'Phase', '3', 'program', 'for', 'its', 'once-daily', 'Janus', 'kina
se', '1', '(', 'JAK1', ')', 'inhibitor', 'PF-04965842', ',', 'to', 'evaluat
e', 'the', 'efficacy', 'and', 'safety', 'of', 'PF-04965842', 'for', 'the',
'treatment', 'of', 'moderate-to-severe', 'atopic', 'dermatitis', '(', 'AD',
')', '.', 'This', 'is', 'the', 'first', 'trial', 'in', 'the', 'JAK1', 'Atopi
c', 'Dermatitis', 'Efficacy', 'and', 'safety', '(', 'JADE', ')', 'global',
'development', 'program', '.', '“', 'By', 'initiating', 'this', 'Phase',
'3', 'program', 'in', 'atopic', 'dermatitis', ',', 'we', 'hope', 'to', 'prov
ide', 'a', 'new', 'potential', 'treatment', 'option', 'for', 'people', 'suff
ering', 'with', 'this', 'condition', ',', '”', 'said', 'Michael', 'Corbo',
',', 'Chief', 'Development', 'Officer', ',', 'Inflammation', '&', 'Immunolog
y', ',', 'Pfizer', 'Global', 'Product', 'Development', '.', '“', 'Pfizer',
'continues', 'to', 'build', 'a', 'leadership', 'position', 'in', 'inflammati
on', 'and', 'immunology', 'research', 'with', 'the', 'advancement', 'of', 't
his', 'important', ',', 'Pfizer-discovered', 'investigational', 'oral', 'JAK
1', 'inhibitor', '.', '”', 'About', 'the', 'Phase', '3', 'Trial', 'B745101
2', 'This', 'Phase', '3', 'trial', 'is', 'a', 'randomized', ',', 'double-bli
nd', ',', 'placebo-controlled', ',', 'parallel-group', 'study', 'designed',
'to', 'evaluate', 'the', 'efficacy', 'and', 'safety', 'of', 'PF-04965842',
'in', '375', 'patients', '12', 'years', 'and', 'older', 'with', 'moderate-to
-severe', 'AD', '.', 'Trial', 'participants', 'will', 'be', 'randomly', 'ass
igned', 'to', 'receive', '200', 'mg', 'or', '100', 'mg', 'once', 'daily', 'o
r', 'placebo', '.', 'The', 'primary', 'endpoints', 'are', 'the', 'proportio
n', 'of', 'patients', 'achieving', 'an', 'Investigator', 'Global', 'Assessme
nt', '(', 'IGA', ')', 'score', 'of', '0/1', 'and', '≥', '2', 'point', 'impro
vement', ',', 'and', 'the', 'proportion', 'of', 'patients', 'with', 'at', 'l
east', 'a', '75', '%', 'or', 'greater', 'change', 'from', 'baseline', 'in',
'their', 'Eczema', 'Area', 'and', 'Severity', 'Index', '(', 'EASI', ')', 'sc
ore', '.', 'Key', 'secondary', 'endpoints', 'include', 'the', 'pruritus', 'n
umerical', 'rating', 'scale', ',', 'the', 'Pruritus', 'and', 'Symptoms', 'As
sessment', 'for', 'Atopic', 'Dermatitis', '(', 'PSAAD', ')', 'electronic',
'diary', 'and', 'safety', 'measures', 'such', 'as', 'the', 'incidence', 'o
f', 'treatment', 'emergent', 'adverse', 'events', 'and', 'laboratory', 'abno
rmalities', '.', 'The', 'treatment', 'duration', 'will', 'be', '12', 'week
s', ',', 'the', 'same', 'duration', 'as', 'the', 'Phase', '2b', 'study', 'B7
451006', ',', 'with', 'a', '4', 'week', 'safety', 'follow-up', 'period', 'o
r', 'the', 'option', 'to', 'enter', 'a', 'long-term', 'extension', 'study',
'(', 'B7451015', ')', 'at', 'Week', '12', '.', 'More', 'on', 'the', 'study',
'can', 'be', 'found', 'on', 'www.clinicaltrials.gov', 'under', 'the', 'ident
ifier', 'NCT03349060', '.', 'The', 'design', 'of', 'the', 'Phase', '3', 'tri
al', 'is', 'based', 'on', 'the', 'Phase', '2', 'results', 'that', 'were', 'p
resented', 'at', 'the', '26th', 'Congress', 'of', 'the', 'European', 'Academ
y', 'of', 'Dermatology', 'and', 'Venereology', 'in', 'September', '2017',
'.', 'About', 'Atopic', 'Dermatitis', 'Atopic', 'dermatitis', ',', 'also',
'commonly', 'called', 'atopic', 'eczema', ',', 'is', 'inflammation', 'of',
'the', 'skin', 'and', 'characterized', 'by', 'erythema', '(', 'redness',
')', ',', 'itching', '(', 'pruritus', ')', ',', 'induration', '(', 'hardenin
g', ')', '/', 'papulation', '(', 'formation', 'of', 'papules', ')', ',', 'an
d', 'oozing/crusting', '.', 'About', 'Pfizer', '’s', 'Kinase', 'Inhibitor',
'Leadership', 'PF-04965842', 'is', 'an', 'oral', 'small', 'molecule', 'tha
t', 'selectively', 'inhibits', 'Janus', 'kinase', '1', '(', 'JAK1', ')',
```

```
'.', 'Inhibition', 'of', 'JAK1', 'modulates', 'multiple', 'cytokines', 'invo
lved', 'in', 'pathophysiology', 'of', 'AD', 'including', 'interleukin', '(',
'IL', ')', '-4', ',', 'IL-13', ',', 'IL-31', 'and', 'interferon', 'gamma',
'(', 'IFNɣ', ')', '.', '“', 'We', 'look', 'forward', 'to', 'advancing', 'oth
er', 'kinase', 'inhibitors', 'currently', 'in', 'mid-stage', 'research', 'fo
r', 'other', 'diseases', 'such', 'as', 'alopecia', ',', 'psoriasis', ',', 'i
nflammatory', 'bowel', 'disease', ',', 'and', 'rheumatoid', 'arthritis',
',', '”', 'said', 'Michael', 'Vincent', ',', 'M.D', ',', 'Ph.D.', ',', 'Seni
or', 'Vice', 'President', 'and', 'Chief', 'Scientific', 'Officer', 'of', 'In
flammation', 'and', 'Immunology', 'Research', 'at', 'Pfizer', '.', 'Pfizer',
'has', 'established', 'a', 'leading', 'kinase', 'research', 'capability', 'w
ith', 'multiple', 'unique', 'kinase', 'inhibitor', 'therapies', 'in', 'devel
opment', '.', 'As', 'a', 'pioneer', 'in', 'JAK', 'science', ',', 'the', 'Com
pany', 'is', 'advancing', 'several', 'investigational', 'programs', 'with',
'novel', 'selectivity', 'profiles', ',', 'which', ',', 'if', 'successful',
',', 'could', 'potentially', 'deliver', 'transformative', 'therapies', 'fo
r', 'patients', '.', 'Pfizer', 'has', 'three', 'additional', 'kinase', 'inhi
bitors', 'in', 'Phase', '2', 'development', 'across', 'multiple', 'indicatio
ns', ':', 'PF-06651600', ':', 'A', 'JAK3', 'inhibitor', 'under', 'investigat
ion', 'for', 'the', 'treatment', 'of', 'rheumatoid', 'arthritis', ',', 'ulce
rative', 'colitis', 'and', 'alopecia', 'areata', 'PF-06700841', ':', 'A', 't
yrosine', 'kinase', '2', '(', 'TYK2', ')', '/', 'JAK1', 'inhibitor', 'unde
r', 'investigation', 'for', 'the', 'treatment', 'of', 'psoriasis', ',', 'ulc
erative', 'colitis', 'and', 'alopecia', 'areata', 'PF-06650833', ':', 'An',
'interleukin-1', 'receptor-associated', 'kinase', '4', '(', 'IRAK4', ')', 'i
nhibitor', 'under', 'investigation', 'for', 'the', 'treatment', 'of', 'rheum
atoid', 'arthritis', 'Working', 'together', 'for', 'a', 'healthier', 'worl
d', '®', 'At', 'Pfizer', ',', 'we', 'apply', 'science', 'and', 'our', 'globa
l', 'resources', 'to', 'bring', 'therapies', 'to', 'people', 'that', 'exten
d', 'and', 'significantly', 'improve', 'their', 'lives', '.', 'We', 'striv
e', 'to', 'set', 'the', 'standard', 'for', 'quality', ',', 'safety', 'and',
'value', 'in', 'the', 'discovery', ',', 'development', 'and', 'manufacture',
'of', 'health', 'care', 'products', '.', 'Our', 'global', 'portfolio', 'incl
udes', 'medicines', 'and', 'vaccines', 'as', 'well', 'as', 'many', 'of', 'th
e', 'world', "'s", 'best-known', 'consumer', 'health', 'care', 'products',
'.', 'Every', 'day', ',', 'Pfizer', 'colleagues', 'work', 'across', 'develop
ed', 'and', 'emerging', 'markets', 'to', 'advance', 'wellness', ',', 'preven
tion', ',', 'treatments', 'and', 'cures', 'that', 'challenge', 'the', 'mos
t', 'feared', 'diseases', 'of', 'our', 'time', '.', 'Consistent', 'with', 'o
ur', 'responsibility', 'as', 'one', 'of', 'the', 'world', "'s", 'premier',
'innovative', 'biopharmaceutical', 'companies', ',', 'we', 'collaborate', 'w
ith', 'health', 'care', 'providers', ',', 'governments', 'and', 'local', 'co
mmunities', 'to', 'support', 'and', 'expand', 'access', 'to', 'reliable',
',', 'affordable', 'health', 'care', 'around', 'the', 'world', '.', 'For',
'more', 'than', '150', 'years', ',', 'we', 'have', 'worked', 'to', 'make',
'a', 'difference', 'for', 'all', 'who', 'rely', 'on', 'us', '.', 'We', 'rout
inely', 'post', 'information', 'that', 'may', 'be', 'important', 'to', 'inve
stors', 'on', 'our', 'website', 'at', 'www.pfizer.com', '.', 'In', 'additio
n', ',', 'to', 'learn', 'more', ',', 'please', 'visit', 'us', 'on', 'www.pfi
zer.com', 'and', 'follow', 'us', 'on', 'Twitter', 'at', '@Pfizer', 'and', '@
Pfizer_News', ',', 'LinkedIn', ',', 'YouTube', 'and', 'like', 'us', 'on', 'F
acebook', 'at', 'Facebook.com/Pfizer', '.', 'DISCLOSURE', 'NOTICE', ':', 'Th
e', 'information', 'contained', 'in', 'this', 'release', 'is', 'as', 'of',
'December', '12', ',', '2017', '.', 'Pfizer', 'assumes', 'no', 'obligation',
'to', 'update', 'forward-looking', 'statements', 'contained', 'in', 'this',
'release', 'as', 'the', 'result', 'of', 'new', 'information', 'or', 'futur
e', 'events', 'or', 'developments', '.', 'This', 'release', 'contains', 'for
ward-looking', 'information', 'about', 'PF-04965842', 'and', 'Pfizer', '’s',
'ongoing', 'investigational', 'programs', 'in', 'kinase', 'inhibitor', 'ther
apies', ',', 'including', 'their', 'potential', 'benefits', ',', 'that', 'in
volves', 'substantial', 'risks', 'and', 'uncertainties', 'that', 'could', 'c
```

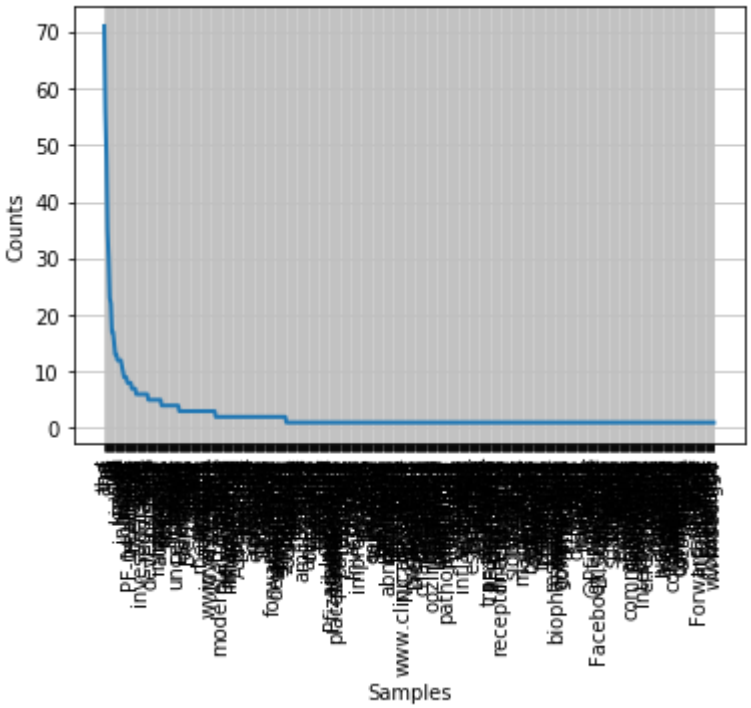
```
ause', 'actual', 'results', 'to', 'differ', 'materially', 'from', 'those',
'expressed', 'or', 'implied', 'by', 'such', 'statements', '.', 'Risks', 'an
d', 'uncertainties', 'include', ',', 'among', 'other', 'things', ',', 'the',
'uncertainties', 'inherent', 'in', 'research', 'and', 'development', ',', 'i
ncluding', 'the', 'ability', 'to', 'meet', 'anticipated', 'clinical', 'tria
l', 'commencement', 'and', 'completion', 'dates', 'and', 'regulatory', 'subm
ission', 'dates', ',', 'as', 'well', 'as', 'the', 'possibility', 'of', 'unfa
vorable', 'clinical', 'trial', 'results', ',', 'including', 'unfavorable',
'new', 'clinical', 'data', 'and', 'additional', 'analyses', 'of', 'existin
g', 'data', ';', 'risks', 'associated', 'with', 'preliminary', 'data', ';',
'the', 'risk', 'that', 'clinical', 'trial', 'data', 'are', 'subject', 'to',
'differing', 'interpretations', ',', 'and', ',', 'even', 'when', 'we', 'vie
w', 'data', 'as', 'sufficient', 'to', 'support', 'the', 'safety', 'and/or',
'effectiveness', 'of', 'a', 'product', 'candidate', ',', 'regulatory', 'auth
orities', 'may', 'not', 'share', 'our', 'views', 'and', 'may', 'require', 'a
dditional', 'data', 'or', 'may', 'deny', 'approval', 'altogether', ';', 'whe
ther', 'regulatory', 'authorities', 'will', 'be', 'satisfied', 'with', 'th
e', 'design', 'of', 'and', 'results', 'from', 'our', 'clinical', 'studies',
';', 'whether', 'and', 'when', 'drug', 'applications', 'may', 'be', 'filed',
'in', 'any', 'jurisdictions', 'for', 'any', 'potential', 'indication', 'fo
r', 'PF-04965842', 'or', 'any', 'other', 'investigational', 'kinase', 'inhib
itor', 'therapies', ';', 'whether', 'and', 'when', 'any', 'such', 'applicati
ons', 'may', 'be', 'approved', 'by', 'regulatory', 'authorities', ',', 'whic
h', 'will', 'depend', 'on', 'the', 'assessment', 'by', 'such', 'regulatory',
'authorities', 'of', 'the', 'benefit-risk', 'profile', 'suggested', 'by', 't
he', 'totality', 'of', 'the', 'efficacy', 'and', 'safety', 'information', 's
ubmitted', ',', 'and', ',', 'if', 'approved', ',', 'whether', 'PF-04965842',
'or', 'any', 'such', 'other', 'investigational', 'kinase', 'inhibitor', 'the
rapies', 'will', 'be', 'commercially', 'successful', ';', 'decisions', 'by',
'regulatory', 'authorities', 'regarding', 'labeling', ',', 'safety', 'and',
'other', 'matters', 'that', 'could', 'affect', 'the', 'availability', 'or',
'commercial', 'potential', 'of', 'PF-04965842', 'or', 'any', 'other', 'inves
tigational', 'kinase', 'inhibitor', 'therapies', ';', 'and', 'competitive',
'developments', '.', 'A', 'further', 'description', 'of', 'risks', 'and', 'u
ncertainties', 'can', 'be', 'found', 'in', 'Pfizer', '’s', 'Annual', 'Repor
t', 'on', 'Form', '10-K', 'for', 'the', 'fiscal', 'year', 'ended', 'Decembe
r', '31', ',', '2016', 'and', 'in', 'its', 'subsequent', 'reports', 'on', 'F
orm', '10-Q', ',', 'including', 'in', 'the', 'sections', 'thereof', 'caption
ed', '“', 'Risk', 'Factors', '”', 'and', '“', 'Forward-Looking', 'Informatio
n', 'and', 'Factors', 'That', 'May', 'Affect', 'Future', 'Results', '”',
',', 'as', 'well', 'as', 'in', 'its', 'subsequent', 'reports', 'on', 'Form',
'8-K', ',', 'all', 'of', 'which', 'are', 'filed', 'with', 'the', 'U.S.', 'Se
curities', 'and', 'Exchange', 'Commission', 'and', 'available', 'at', 'www.s
ec.gov', 'and', 'www.pfizer.com', '.']
```

In [36]:

```
freqDist = nltk.FreqDist(st_tokens_lst)
```

In [37]:

```
freqDist.plot()
```



In [38]:

```
from collections import Counter
token_counts = Counter(st_tokens_lst)
print(token_counts)
```

```
Counter({',': 71, 'and': 56, 'the': 48, 'of': 35, '.': 30, 'in': 23, 'to': 2
2, '(': 17, ')': 17, 'for': 15, 'with': 13, 'as': 13, 'a': 12, 'kinase': 12,
'or': 12, 'on': 12, 'Pfizer': 11, 'inhibitor': 10, 'safety': 9, 'be': 9, 'th
at': 9, ';': 8, 'Phase': 8, 'PF-04965842': 8, 'at': 8, 'treatment': 7, 'is':
7, 'other': 7, 'therapies': 7, ':': 6, 'JAK1': 6, 'trial': 6, 'investigation
al': 6, 'such': 6, 'by': 6, 'our': 6, 'may': 6, 'regulatory': 6, 'data': 6,
'any': 6, 'study': 5, '3': 5, 'development': 5, '“': 5, 'this': 5, 'we': 5,
'”': 5, 'will': 5, 'including': 5, 'information': 5, 'clinical': 5, 'authori
ties': 5, 'program': 4, 'Atopic': 4, 'potential': 4, 'research': 4, 'patient
s': 4, '12': 4, 'The': 4, '2': 4, 'under': 4, 'results': 4, 'world': 4, 'hea
lth': 4, 'care': 4, 'us': 4, 'uncertainties': 4, 'whether': 4, 'Global': 3,
'December': 3, '2017': 3, 'its': 3, 'efficacy': 3, 'atopic': 3, 'dermatiti
s': 3, 'AD': 3, 'This': 3, 'Dermatitis': 3, 'global': 3, 'new': 3, 'About':
3, 'are': 3, 'from': 3, 'their': 3, '’s': 3, 'multiple': 3, 'We': 3, 'alopec
ia': 3, 'rheumatoid': 3, 'arthritis': 3, 'which': 3, 'could': 3, 'additiona
l': 3, 'A': 3, 'investigation': 3, 'well': 3, 'www.pfizer.com': 3, 'releas
e': 3, 'risks': 3, 'when': 3, 'Form': 3, 'B7451012': 2, 'Janus': 2, '1': 2,
'evaluate': 2, 'moderate-to-severe': 2, 'option': 2, 'people': 2, 'said': 2,
'Michael': 2, 'Chief': 2, 'Development': 2, 'Officer': 2, 'Inflammation': 2,
'Immunology': 2, 'inflammation': 2, 'important': 2, 'oral': 2, 'Trial': 2,
'years': 2, 'mg': 2, 'endpoints': 2, 'proportion': 2, 'an': 2, 'Assessment':
2, 'score': 2, 'include': 2, 'pruritus': 2, 'events': 2, 'duration': 2, '4':
2, 'can': 2, 'found': 2, 'design': 2, '/': 2, 'advancing': 2, 'inhibitors':
2, 'diseases': 2, 'psoriasis': 2, 'has': 2, 'science': 2, 'programs': 2, 'i
f': 2, 'successful': 2, 'across': 2, 'ulcerative': 2, 'colitis': 2, 'areat
a': 2, 'products': 2, "'s": 2, 'support': 2, 'more': 2, 'all': 2, 'containe
d': 2, 'forward-looking': 2, 'statements': 2, 'developments': 2, 'dates': 2,
'unfavorable': 2, 'applications': 2, 'filed': 2, 'approved': 2, 'subsequen
t': 2, 'reports': 2, 'Factors': 2, '2017-12-14': 1, '00:00:00': 1, 'commenc
e': 1, 'pivotal': 1, 'North': 1, 'America': 1, 'Australia': 1, 'Europe': 1,
'broader': 1, 'regional': 1, 'rollout': 1, '2018': 1, 'Thursday': 1, '14':
1, '-': 1, '7:30': 1, 'amESTPfizer': 1, 'Inc.': 1, 'NYSE': 1, 'PFE': 1, 'tod
ay': 1, 'announced': 1, 'initiation': 1, 'once-daily': 1, 'first': 1, 'Effic
acy': 1, 'JADE': 1, 'By': 1, 'initiating': 1, 'hope': 1, 'provide': 1, 'suff
ering': 1, 'condition': 1, 'Corbo': 1, '&': 1, 'Product': 1, 'continues': 1,
'build': 1, 'leadership': 1, 'position': 1, 'immunology': 1, 'advancement':
1, 'Pfizer-discovered': 1, 'randomized': 1, 'double-blind': 1, 'placebo-cont
rolled': 1, 'parallel-group': 1, 'designed': 1, '375': 1, 'older': 1, 'parti
cipants': 1, 'randomly': 1, 'assigned': 1, 'receive': 1, '200': 1, '100': 1,
'once': 1, 'daily': 1, 'placebo': 1, 'primary': 1, 'achieving': 1, 'Investig
ator': 1, 'IGA': 1, '0/1': 1, '≥': 1, 'point': 1, 'improvement': 1, 'least':
1, '75': 1, '%': 1, 'greater': 1, 'change': 1, 'baseline': 1, 'Eczema': 1,
'Area': 1, 'Severity': 1, 'Index': 1, 'EASI': 1, 'Key': 1, 'secondary': 1,
'numerical': 1, 'rating': 1, 'scale': 1, 'Pruritus': 1, 'Symptoms': 1, 'PSAA
D': 1, 'electronic': 1, 'diary': 1, 'measures': 1, 'incidence': 1, 'emergen
t': 1, 'adverse': 1, 'laboratory': 1, 'abnormalities': 1, 'weeks': 1, 'sam
e': 1, '2b': 1, 'B7451006': 1, 'week': 1, 'follow-up': 1, 'period': 1, 'ente
r': 1, 'long-term': 1, 'extension': 1, 'B7451015': 1, 'Week': 1, 'More': 1,
'www.clinicaltrials.gov': 1, 'identifier': 1, 'NCT03349060': 1, 'based': 1,
'were': 1, 'presented': 1, '26th': 1, 'Congress': 1, 'European': 1, 'Academ
y': 1, 'Dermatology': 1, 'Venereology': 1, 'September': 1, 'also': 1, 'commo
nly': 1, 'called': 1, 'eczema': 1, 'skin': 1, 'characterized': 1, 'erythem
a': 1, 'redness': 1, 'itching': 1, 'induration': 1, 'hardening': 1, 'papulat
ion': 1, 'formation': 1, 'papules': 1, 'oozing/crusting': 1, 'Kinase': 1, 'I
nhibitor': 1, 'Leadership': 1, 'small': 1, 'molecule': 1, 'selectively': 1,
```

```
'inhibits': 1, 'Inhibition': 1, 'modulates': 1, 'cytokines': 1, 'involved':
1, 'pathophysiology': 1, 'interleukin': 1, 'IL': 1, '-4': 1, 'IL-13': 1, 'IL
-31': 1, 'interferon': 1, 'gamma': 1, 'IFNɣ': 1, 'look': 1, 'forward': 1, 'c
urrently': 1, 'mid-stage': 1, 'inflammatory': 1, 'bowel': 1, 'disease': 1,
'Vincent': 1, 'M.D': 1, 'Ph.D.': 1, 'Senior': 1, 'Vice': 1, 'President': 1,
'Scientific': 1, 'Research': 1, 'established': 1, 'leading': 1, 'capabilit
y': 1, 'unique': 1, 'As': 1, 'pioneer': 1, 'JAK': 1, 'Company': 1, 'severa
l': 1, 'novel': 1, 'selectivity': 1, 'profiles': 1, 'potentially': 1, 'deliv
er': 1, 'transformative': 1, 'three': 1, 'indications': 1, 'PF-06651600': 1,
'JAK3': 1, 'PF-06700841': 1, 'tyrosine': 1, 'TYK2': 1, 'PF-06650833': 1, 'A
n': 1, 'interleukin-1': 1, 'receptor-associated': 1, 'IRAK4': 1, 'Working':
1, 'together': 1, 'healthier': 1, '®': 1, 'At': 1, 'apply': 1, 'resources':
1, 'bring': 1, 'extend': 1, 'significantly': 1, 'improve': 1, 'lives': 1, 's
trive': 1, 'set': 1, 'standard': 1, 'quality': 1, 'value': 1, 'discovery':
1, 'manufacture': 1, 'Our': 1, 'portfolio': 1, 'includes': 1, 'medicines':
1, 'vaccines': 1, 'many': 1, 'best-known': 1, 'consumer': 1, 'Every': 1, 'da
y': 1, 'colleagues': 1, 'work': 1, 'developed': 1, 'emerging': 1, 'markets':
1, 'advance': 1, 'wellness': 1, 'prevention': 1, 'treatments': 1, 'cures':
1, 'challenge': 1, 'most': 1, 'feared': 1, 'time': 1, 'Consistent': 1, 'resp
onsibility': 1, 'one': 1, 'premier': 1, 'innovative': 1, 'biopharmaceutica
l': 1, 'companies': 1, 'collaborate': 1, 'providers': 1, 'governments': 1,
'local': 1, 'communities': 1, 'expand': 1, 'access': 1, 'reliable': 1, 'affo
rdable': 1, 'around': 1, 'For': 1, 'than': 1, '150': 1, 'have': 1, 'worked':
1, 'make': 1, 'difference': 1, 'who': 1, 'rely': 1, 'routinely': 1, 'post':
1, 'investors': 1, 'website': 1, 'In': 1, 'addition': 1, 'learn': 1, 'pleas
e': 1, 'visit': 1, 'follow': 1, 'Twitter': 1, '@Pfizer': 1, '@Pfizer_News':
1, 'LinkedIn': 1, 'YouTube': 1, 'like': 1, 'Facebook': 1, 'Facebook.com/Pfiz
er': 1, 'DISCLOSURE': 1, 'NOTICE': 1, 'assumes': 1, 'no': 1, 'obligation':
1, 'update': 1, 'result': 1, 'future': 1, 'contains': 1, 'about': 1, 'ongoin
g': 1, 'benefits': 1, 'involves': 1, 'substantial': 1, 'cause': 1, 'actual':
1, 'differ': 1, 'materially': 1, 'those': 1, 'expressed': 1, 'implied': 1,
'Risks': 1, 'among': 1, 'things': 1, 'inherent': 1, 'ability': 1, 'meet': 1,
'anticipated': 1, 'commencement': 1, 'completion': 1, 'submission': 1, 'poss
ibility': 1, 'analyses': 1, 'existing': 1, 'associated': 1, 'preliminary':
1, 'risk': 1, 'subject': 1, 'differing': 1, 'interpretations': 1, 'even': 1,
'view': 1, 'sufficient': 1, 'and/or': 1, 'effectiveness': 1, 'product': 1,
'candidate': 1, 'not': 1, 'share': 1, 'views': 1, 'require': 1, 'deny': 1,
'approval': 1, 'altogether': 1, 'satisfied': 1, 'studies': 1, 'drug': 1, 'ju
risdictions': 1, 'indication': 1, 'depend': 1, 'assessment': 1, 'benefit-ris
k': 1, 'profile': 1, 'suggested': 1, 'totality': 1, 'submitted': 1, 'commerc
ially': 1, 'decisions': 1, 'regarding': 1, 'labeling': 1, 'matters': 1, 'aff
ect': 1, 'availability': 1, 'commercial': 1, 'competitive': 1, 'further': 1,
'description': 1, 'Annual': 1, 'Report': 1, '10-K': 1, 'fiscal': 1, 'year':
1, 'ended': 1, '31': 1, '2016': 1, '10-Q': 1, 'sections': 1, 'thereof': 1,
'captioned': 1, 'Risk': 1, 'Forward-Looking': 1, 'Information': 1, 'That':
1, 'May': 1, 'Affect': 1, 'Future': 1, 'Results': 1, '8-K': 1, 'U.S.': 1, 'S
ecurities': 1, 'Exchange': 1, 'Commission': 1, 'available': 1, 'www.sec.go
v': 1})
```

In [39]:

```
def dict_with_frequency(low, high):
    freq_words = {x:token_counts[x] for x in token_counts if token_counts[x] > low and toke
    return freq_words
```

High frequency words

In [40]:

```
high_freq_words = dict_with_frequency(6, 100)
print(high_freq_words)
```

```
{'to': 22, 'with': 13, 'in': 23, ',': 71, 'and': 56, ';': 8, '(': 17, ')': 1
7, 'the': 48, 'of': 35, 'a': 12, 'Phase': 8, 'for': 15, 'kinase': 12, 'inhib
itor': 10, 'PF-04965842': 8, 'safety': 9, 'treatment': 7, '.': 30, 'is': 7,
'Pfizer': 11, 'be': 9, 'or': 12, 'at': 8, 'as': 13, 'on': 12, 'that': 9, 'ot
her': 7, 'therapies': 7}
```

Mid frequency words

In [41]:

```
mid_freq_words = dict_with_frequency(2, 7)
print(mid_freq_words)
```

```
{'Global': 3, 'program': 4, 'study': 5, 'December': 3, '2017': 3, ':': 6,
'3': 5, 'its': 3, 'JAK1': 6, 'efficacy': 3, 'atopic': 3, 'dermatitis': 3, 'A
D': 3, 'This': 3, 'trial': 6, 'Atopic': 4, 'Dermatitis': 3, 'global': 3, 'de
velopment': 5, '“': 5, 'this': 5, 'we': 5, 'new': 3, 'potential': 4, '”': 5,
'research': 4, 'investigational': 6, 'About': 3, 'patients': 4, '12': 4, 'wi
ll': 5, 'The': 4, 'are': 3, '2': 4, 'from': 3, 'their': 3, 'such': 6, 'unde
r': 4, 'results': 4, 'by': 6, '’s': 3, 'multiple': 3, 'including': 5, 'We':
3, 'alopecia': 3, 'rheumatoid': 3, 'arthritis': 3, 'which': 3, 'could': 3,
'additional': 3, 'A': 3, 'investigation': 3, 'world': 4, 'our': 6, 'health':
4, 'care': 4, 'well': 3, 'us': 4, 'information': 5, 'may': 6, 'www.pfizer.co
m': 3, 'release': 3, 'risks': 3, 'uncertainties': 4, 'clinical': 5, 'regulat
ory': 6, 'data': 6, 'when': 3, 'authorities': 5, 'whether': 4, 'any': 6, 'Fo
rm': 3}
```

Low frequency words

In [42]:

```
low_freq_words = dict_with_frequency(0, 3)
print(low_freq_words)
```

```
{'2017-12-14': 1, '00:00:00': 1, 'commence': 1, 'pivotal': 1, 'B7451012':
2, 'North': 1, 'America': 1, 'Australia': 1, 'Europe': 1, 'broader': 1, 'r
egional': 1, 'rollout': 1, '2018': 1, 'Thursday': 1, '14': 1, '-': 1, '7:3
0': 1, 'amESTPfizer': 1, 'Inc.': 1, 'NYSE': 1, 'PFE': 1, 'today': 1, 'anno
unced': 1, 'initiation': 1, 'once-daily': 1, 'Janus': 2, '1': 2, 'evaluat
e': 2, 'moderate-to-severe': 2, 'first': 1, 'Efficacy': 1, 'JADE': 1, 'B
y': 1, 'initiating': 1, 'hope': 1, 'provide': 1, 'option': 2, 'people': 2,
'suffering': 1, 'condition': 1, 'said': 2, 'Michael': 2, 'Corbo': 1, 'Chie
f': 2, 'Development': 2, 'Officer': 2, 'Inflammation': 2, '&': 1, 'Immunol
ogy': 2, 'Product': 1, 'continues': 1, 'build': 1, 'leadership': 1, 'posit
ion': 1, 'inflammation': 2, 'immunology': 1, 'advancement': 1, 'importan
t': 2, 'Pfizer-discovered': 1, 'oral': 2, 'Trial': 2, 'randomized': 1, 'do
uble-blind': 1, 'placebo-controlled': 1, 'parallel-group': 1, 'designed':
1, '375': 1, 'years': 2, 'older': 1, 'participants': 1, 'randomly': 1, 'as
signed': 1, 'receive': 1, '200': 1, 'mg': 2, '100': 1, 'once': 1, 'daily':
1, 'placebo': 1, 'primary': 1, 'endpoints': 2, 'proportion': 2, 'achievin
g': 1, 'an': 2, 'Investigator': 1, 'Assessment': 2, 'IGA': 1, 'score': 2,
'0/1': 1, '≥': 1, 'point': 1, 'improvement': 1, 'least': 1, '75': 1, '%':
1, 'greater': 1, 'change': 1, 'baseline': 1, 'Eczema': 1, 'Area': 1, 'Seve
```

In [43]:

```
from nltk.tokenize import TreebankWordTokenizer
from nltk.tokenize import PunktSentenceTokenizer
```

Text Normalization : Find root words with stemming or lemmatization to reduce number of features. Case conversion (casing) and acronyms

- **Stemming/Lemmatization -** treat stop words OR high frequency small words. Map single/multiple tokens with terms. Example map window - window, windows; de-accented term - resume.
  - **Stemming** algorithms work by cutting off the end of the word, and in some cases also the beginning while looking for the root. This indiscriminate cutting can be successful in some occasions, but not always, that is why we affirm that this is an approach that has some limitations.
    - The best-known and most popular stemming approach for English is the Porter stemming algorithm, also known as the Porter stemmer. It is a collection of rules (or, if you prefer, heuristics) designed to reflect how English handles inflections. **For example, the Porter stemmer chops both apple and apples down to appl, and it stems berry and berries to berri.**
    - Porter Stemmer (5 phase stemming), Lovins Stemmer (1 phase stemming). On English, stemming is not a good idea to convert tokens to terms (recall will be high, precision is going to be low). This technique works well with Spanish, Finish, German, etc., language.
      - Phase I rules :
        - SSES -> SS : caresses -> caress
        - IES -> I : ponies -> poni
  - **Lemmatization** on the other hand takes into consideration the morphological analysis of the words. To do so it is necessary to have detailed dictionaries the algorithm can look back at to link the form back to its lemma.
    - Lemmatization does not simply chop off inflections, but instead relies on a lexical knowledge base like , WordNet (https://wordnet.princeton.edu/) to obtain the correct base forms of words.
    - For example, WordNet lemmatizes **geese** to **goose**, **meanness to meanness** and **meaning to meaning**. In these examples, it outperforms than the Porter stemmer.

- **Stemming VS Lemmatization :** Lemmatization or Stemming has limits. For example, Porter stems both happiness and happy to happi, while WordNet lemmatizes the two words to themselves. The WordNet lemmatizer also requires specifying the word's part of speech—otherwise, it assumes the word is a noun. Finally, lemmatization cannot handle unknown words: for example, Porter stems both iphone and iphones to iphon, while WordNet lemmatizes both words to themselves. In general, lemmatization offers better precision than stemming, but at the expense of recall.

- Case folding - reduce all letters to lowercase

In [2]:

```
import nltk
```

In [4]:

```
#nltk.download()
```

In [1]:

```
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
```

In [2]:

```
print(type(stopwords))
```

```
<class 'nltk.corpus.util.LazyCorpusLoader'>
```

In [4]:

```
[x for x in stopwords.words('english')]
```

Out[4]:

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
```

In [46]:

```
corpus = []

#sample_text = re.sub('[^\w\d\-\:]', ' ', sample_text) #substitute non alphabets with space
ps = PorterStemmer()

words = [ps.stem(word) for word in sample_text if not word in set(stopwords.words('english'
text_corpus = ''.join(words)
corpus.append(text_corpus)
```

In [47]:

```
corpus
```

Out[47]:

["2017-12-14 00:00:00 Glbl prgr cence wh pvl u B7451012 n Nrh Aerc, Aurl n Eurpe; brer regnl rllu n 2018 Thur, Deceber 14, 2017 - 7:30ESTPfzer Inc. (NY SE:PFE) nnunce he nn f Phe 3 prgr fr nce-l Jnu kne 1 (JAK1) nhbr PF-04965 842, evlue he effcc n fe f PF-04965842 fr he reen f ere--evere pc er (AD). Th he fr rl n he JAK1 Apc Der Effcc n fe (JADE) glbl evelpen prgr. “B nng h Phe 3 prgr n pc er, we hpe prve new penl reen pn fr peple ufferng wh h cn n,” Mchel Crb, Chef Develpen Offcer, Infln & Iunlg, Pfzer Glbl Pruc Develpe n. “Pfzer cnnue bul leerhp pn n nfln n unlg reerch wh he vnceen f h prn, P fzer-cvere nvegnl rl JAK1 nhbr.” Abu he Phe 3 Trl B7451012 Th Phe 3 rl rnz e, uble-bln, plceb-cnrlle, prllel-grup u egne evlue he effcc n fe f PF-0496 5842 n 375 pen 12 er n ler wh ere--evere AD. Trl prcpn wll be rnl gne recev e 200 g r 100 g nce l r plceb. The prr enpn re he prprn f pen chevng n Inveg r Glbl Aeen (IGA) cre f 0/1 n ≥2 pn prveen, n he prprn f pen wh le 75% r g reer chnge fr belne n her Ecze Are n Sever Inex (EASI) cre. Ke ecnr enpn ncl ue he pruru nuercl rng cle, he Pruru n Sp Aeen fr Apc Der (PSAAD) elecrnc r n fe eure uch he ncence f reen eergen vere even n lbrr bnrle. The reen urn wll be 12 week, he e urn he Phe 2b u B7451006, wh 4 week fe fllw-up per r he pn ener lng-er exenn u (B7451015) Week 12. Mre n he u cn be fun n www. clnclrl.gv uner he enfer NCT03349060. The egn f he Phe 3 rl be n he Phe 2 r eul h were preene he 26h Cngre f he Eurpen Ace f Derlg n Venerelg n Sepeber 2017. Abu Apc Der Apc er, l cnl clle pc ecze, nfln f he kn n chrcerze b erh e (rene), chng (pruru), nurn (hrenng)/ppuln (frn f ppule), n zng/crung. Abu Pfzer’ Kne Inhbr Leerhp PF-04965842 n rl ll lecule h elecvel nhb Jnu kne 1 (JAK1). Inhbn f JAK1 ule ulple ckne nvlve n phphlg f AD nclung nerleukn (IL) -4, IL-13, IL-31 n nerfern g (IFNɣ). “We lk frwr vncng her kne nhbr currenl n -ge reerch fr her ee uch lpec, pr, nflr bwel ee, n rheu rhr,” Mchel Vnce n, M.D, Ph.D., Senr Vce Preen n Chef Scenfc Offcer f Infln n Iunlg Reerch P fzer. Pfzer h eblhe leng kne reerch cpbl wh ulple unque kne nhbr herpe n ev elpen. A pneer n JAK cence, he Cpn vncng everl nvegnl prgr wh nvel elecv p rfle, whch, f ucceful, cul penll elver rnfrve herpe fr pen. Pfzer h hree nl kne nhbr n Phe 2 evelpen cr ulple ncn: PF-06651600: A JAK3 nhbr uner nveg n fr he reen f rheu rhr, ulcerve cl n lpec re PF-06700841: A rne kne 2 (T YK2)/JAK1 nhbr uner nvegn fr he reen f pr, ulcerve cl n lpec re PF-066508 33: An nerleukn-1 recepr-ce kne 4 (IRAK4) nhbr uner nvegn fr he reen f rheu rhr Wrkng geher fr helher wrl® A Pfzer, we ppl cence n ur glbl reurce brng herpe peple h exen n gnfcnl prve her lve. We rve e he nr fr qul, fe n vlue n he cver, evelpen n nufcure f helh cre pruc. Our glbl prfl nclue ecne n vcc ne well n f he wrl' be-knwn cnuer helh cre pruc. Ever , Pfzer cllegue wrk cr evelpe n eergng rke vnce wellne, prevenn, reen n cure h chllenge he fer e ee f ur e. Cnen wh ur repnbl ne f he wrl' preer nnvve bphrceucl cpne, we cllbre wh helh cre prver, gvernen n lcl cune uppr n expn cce relble, ffrbl e helh cre run he wrl. Fr re hn 150 er, we hve wrke ke fference fr ll wh r el n u. We runel p nfrn h be prn nver n ur webe www.pfzer.c. In n, lern re, plee v u n www.pfzer.c n fllw u n Twer @Pfzer n @Pfzer_New, LnkeIn, YuT ube n lke u n Fcebk Fcebk.c/Pfzer. DISCLOSURE NOTICE: The nfrn cnne n h rel ee f Deceber 12, 2017. Pfzer ue n blgn upe frwr-lkng een cnne n h relee he reul f new nfrn r fuure even r evelpen. Th relee cnn frwr-lkng nfrn bu PF -04965842 n Pfzer’ ngng nvegnl prgr n kne nhbr herpe, nclung her penl benef, h nvlve ubnl rk n uncerne h cul cue cul reul ffer erll fr he expree r ple b uch een. Rk n uncerne nclue, ng her hng, he uncerne nheren n reerch n evelpe n, nclung he bl ee ncpe clncl rl cenceen n cplen e n regulr ubn e, well h e pbl f unfvrble clncl rl reul, nclung unfvrble new clncl n nl nle f exng ; rk ce wh prelnr ; he rk h clncl rl re ubjec fferng nerpren, n, even when w e vew uffcen uppr he fe n/r effecvene f pruc cne, regulr uhre n hre ur vew n requre nl r en pprvl lgeher; wheher regulr uhre wll be fe wh he egn

```
f n reul fr ur clncl ue; wheher n when rug pplcn  be fle n n jurcn fr n penl
ncn fr PF-04965842 r n her nvegnl kne nhbr herpe; wheher n when n uch pplcn
be pprve b regulr uhre, whch wll epen n he een b uch regulr uhre f he benef-
rk prfle uggee b he l f he effcc n fe nfrn ube, n, f pprve, wheher PF-049658
42 r n uch her nvegnl kne nhbr herpe wll be cercll ucceful; ecn b regulr uhr
e regrng lbelng, fe n her er h cul ffec he vlbl r cercl penl f PF-04965842 r
n her nvegnl kne nhbr herpe; n cpeve evelpen. A furher ecrpn f rk n uncerne
cn be fun n Pfzer’ Annul Repr n Fr 10-K fr he fcl er ene Deceber 31, 2016 n
n  ubequen repr n Fr 10-Q, nclung n he ecn heref cpne “Rk Fcr” n “Frwr-Lkng
Infrn n Fcr Th M Affec Fuure Reul”,  well  n  ubequen repr n Fr 8-K, ll f wh
ch re fle wh he U.S. Secure n Exchnge Cn n vlble  www.ec.gv n www.pfzer.c.
"]
```

In [48]:

```
from nltk.stem import WordNetLemmatizer
```

In [49]:

```
corpus = []

#sample_text = re.sub('[^\w\d\-\:]', ' ', sample_text) #substitute non alphabets with space
lem = WordNetLemmatizer()
words = [lem.lemmatize(word) for word in sample_text if not word in set(stopwords.words('en
text_corpus = ''.join(words)
corpus.append(text_corpus)
```

In [51]:

```
print(corpus)
```

```
["2017-12-14 00:00:00  Glbl prgr  cence wh pvl u B7451012 n Nrh Aerc, Aurl n
Eurpe; brer regnl rllu n 2018 Thur, Deceber 14, 2017 - 7:30ESTPfzer Inc. (NY
SE:PFE)  nnunce he nn f  Phe 3 prgr fr  nce-l Jnu kne 1 (JAK1) nhbr PF-04965
842,  evlue he effcc n fe f PF-04965842 fr he reen f ere--evere pc er (AD).
Th  he fr rl n he JAK1 Apc Der Effcc n fe (JADE) glbl evelpen prgr. “B nng h
Phe 3 prgr n pc er, we hpe  prve  new penl reen pn fr peple uffernng wh h cn
n,”  Mchel Crb, Chef Develpen Offcer, Infln & Iunlg, Pfzer Glbl Pruc Develpe
n. “Pfzer cnnue  bul  leerhp pn n nfln n unlg reerch wh he vnceen f h prn, P
fzer-cvere nvegnl rl JAK1 nhbr.” Abu he Phe 3 Trl B7451012 Th Phe 3 rl  rnz
e, uble-bln, plceb-cnrlle, prllel-grup u egne  evlue he effcc n fe f PF-0496
5842 n 375 pen 12 er n ler wh ere--evere AD. Trl prcpn wll be rnl gne  recev
e 200 g r 100 g nce l r plceb. The prr enpn re he prprn f pen chevng n Inveg
r Glbl Aeen (IGA) cre f 0/1 n ≥2 pn prveen, n he prprn f pen wh  le  75% r g
reer chnge fr belne n her Ecze Are n Sever Inex (EASI) cre. Ke ecnr enpn ncl
ue he pruru nuercl rng cle, he Pruru n Sp Aeen fr Apc Der (PSAAD) elecrnc r
n fe eure uch  he ncence f reen eergen vere even n lbrr bnrle. The reen urn
wll be 12 week, he e urn  he Phe 2b u B7451006, wh  4 week fe fllw-up per r
he pn  ener  lng-er exenn u (B7451015)  Week 12. Mre n he u cn be fun n www.
clnclrl.gv uner he enfer NCT03349060. The egn f he Phe 3 rl  be n he Phe 2 r
eul h were preene  he 26h Cngre f he Eurpen Ace f Derlg n Venerelg n Sepeber
2017. Abu Apc Der Apc er, l cnl clle pc ecze,  nfln f he kn n chrcerze b erh
e (rene), chng (pruru), nurn (hrenng)/ppuln (frn f ppule), n zng/crung. Abu
Pfzer’ Kne Inhbr Leerhp PF-04965842  n rl ll lecule h elecvel nhb Jnu kne 1
(JAK1). Inhbn f JAK1 ule ulple ckne nvlve n phphlg f AD nclung nerleukn (IL)
-4, IL-13, IL-31 n nerfern g (IFNɣ). “We lk frwr  vncng her kne nhbr currenl
n -ge reerch fr her ee uch  lpec, pr, nflr bwel ee, n rheu rhr,”  Mchel Vnce
n, M.D, Ph.D., Senr Vce Preen n Chef Scenfc Offcer f Infln n Iunlg Reerch  P
fzer. Pfzer h eblhe  leng kne reerch cpbl wh ulple unque kne nhbr herpe n ev
elpen. A  pneer n JAK cence, he Cpn  vncng everl nvegnl prgr wh nvel elecv p
rfle, whch, f ucceful, cul penll elver rnfrve herpe fr pen. Pfzer h hree nl
kne nhbr n Phe 2 evelpen cr ulple ncn:    PF-06651600: A JAK3 nhbr uner nveg
n fr he reen f rheu rhr, ulcerve cl n lpec re    PF-06700841: A rne kne 2 (T
YK2)/JAK1 nhbr uner nvegn fr he reen f pr, ulcerve cl n lpec re    PF-066508
33: An nerleukn-1 recepr-ce kne 4 (IRAK4) nhbr uner nvegn fr he reen f rheu
rhr Wrkng geher fr  helher wrl® A Pfzer, we ppl cence n ur glbl reurce  brng
herpe  peple h exen n gnfcnl prve her lve. We rve  e he nr fr qul, fe n vlue
n he cver, evelpen n nufcure f helh cre pruc. Our glbl prfl nclue ecne n vcc
ne  well  n f he wrl' be-knwn cnuer helh cre pruc. Ever , Pfzer cllegue wrk
cr evelpe n eergng rke  vnce wellne, prevenn, reen n cure h chllenge he  fer
e ee f ur e. Cnen wh ur repnbl  ne f he wrl' preer nnvve bphrceucl cpne, we
cllbre wh helh cre prver, gvernen n lcl cune  uppr n expn cce  relble, ffrbl
e helh cre run he wrl. Fr re hn 150 er, we hve wrke  ke  fference fr ll wh r
el n u. We runel p nfrn h  be prn  nver n ur webe  www.pfzer.c. In n,  lern
re, plee v u n www.pfzer.c n fllw u n Twer  @Pfzer n @Pfzer_New, LnkeIn, YuT
ube n lke u n Fcebk  Fcebk.c/Pfzer. DISCLOSURE NOTICE: The nfrn cnne n h rel
ee  f Deceber 12, 2017. Pfzer ue n blgn  upe frwr-lkng een cnne n h relee
he reul f new nfrn r fuure even r evelpen. Th relee cnn frwr-lkng nfrn bu PF
-04965842 n Pfzer’ ngng nvegnl prgr n kne nhbr herpe, nclung her penl benef,
h nvlve ubnl rk n uncerne h cul cue cul reul  ffer erll fr he expree r ple b
uch een. Rk n uncerne nclue, ng her hng, he uncerne nheren n reerch n evelpe
n, nclung he bl  ee ncpe clncl rl cenceen n cplen e n regulr ubn e,  well  h
e pbl f unfvrble clncl rl reul, nclung unfvrble new clncl  n nl nle f exng ;
rk ce wh prelnr ; he rk h clncl rl  re ubjec  fferng nerpren, n, even when w
e vew  uffcen  uppr he fe n/r effecvene f  pruc cne, regulr uhre  n hre ur
vew n  requre nl  r  en pprvl lgeher; wheher regulr uhre wll be fe wh he egn
f n reul fr ur clncl ue; wheher n when rug pplcn  be fle n n jurcn fr n penl
ncn fr PF-04965842 r n her nvegnl kne nhbr herpe; wheher n when n uch pplcn
```

```
be pprve b regulr uhre, whch wll epen n he een b uch regulr uhre f he benef-
rk prfle uggee b he l f he effcc n fe nfrn ube, n, f pprve, wheher PF-049658
42 r n uch her nvegnl kne nhbr herpe wll be cercll ucceful; ecn b regulr uhr
e regrng lbelng, fe n her er h cul ffec he vlbl r cercl penl f PF-04965842 r
n her nvegnl kne nhbr herpe; n cpeve evelpen. A furher ecrpn f rk n uncerne
cn be fun n Pfzer’ Annul Repr n Fr 10-K fr he fcl er ene Deceber 31, 2016 n
n  ubequen repr n Fr 10-Q, nclung n he ecn heref cpne “Rk Fcr” n “Frwr-Lkng
Infrn n Fcr Th M Affec Fuure Reul”,  well  n  ubequen repr n Fr 8-K, ll f wh
ch re fle wh he U.S. Secure n Exchnge Cn n vlble  www.ec.gv n www.pfzer.c.
"]
```

More on Data Preprocessing - the tasks and activities differ based on scenarios.

- **For Text Summarization Task:**
  - Remove repeated words
  - Remove long sentences with stop words and high frequency short words.
- **For Grammer Correction System:**
  - We should not be removing stop words.
  - Can remove math equations, HTML tags.
  - Need detailed analysis before removing abbreviated words.
- **For Sentiment Analysis:**
  - Remove casual words

Tokenization

- Token : An instance of characters, example if token "Friends" occurred twice in document, there will be two entries.
- Term : An entry in dictionary "friend"
- state-of-the-art, co-education, lower-case : for these three tokens, the terms should be as shown below.
  - state-of-the-art : state-of-the-art
  - co-education : co-education, coeducation
  - lower-case : lower-case, lowercase, lower case
- San Francisco : This should not be splitted in to two tokens based on space. "San Francisco" should be one token. We may have to get this from list of cities.

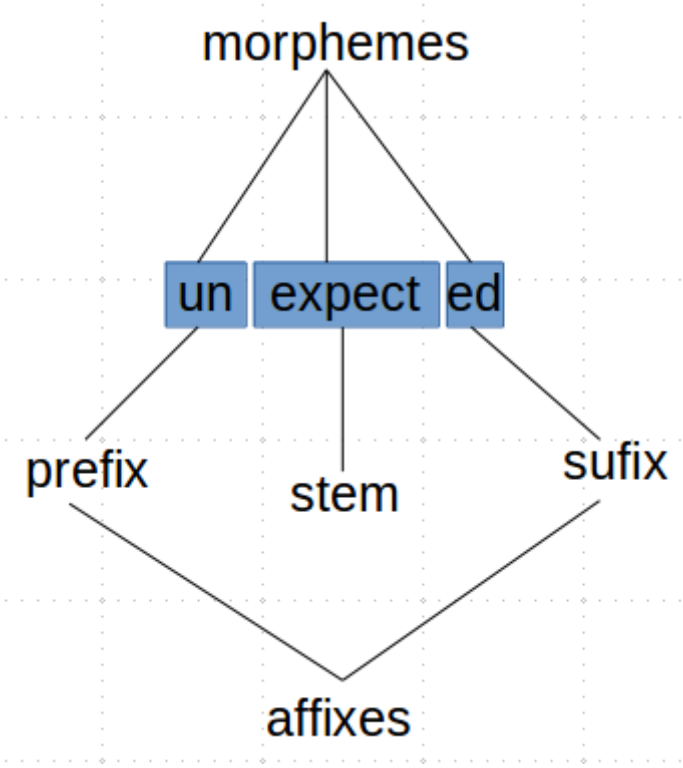
Stop words

- Stop words are the words occur many times in documents. Example the, a, an, to, be, of, to, from, etc.
- **Removing stop words is always not good.** Example King of Denmark - here of is imporant. Flights to London, Flights from London.

More on Lexical and Morphological Analysis:

- **Lexical Analysis:** is the process of breaking down text into meaningful words, phrases. Lexical Analysis is also loosly described as tokenization process.
- **Morphological Analysis:** grammetical analysis of how **words** are formed using morphemes.
  - Morphological analysis is used in word segmentation and POS tagging.
    - **Morphology:** is a branch of linguistics (linguistics - Natural Language analysis using different scientific techniques) that studies how words can be structured and formed.

- **Morphem:** smallest unit in a given language. Morphem may or may not have meaning of its own, but words do have meaning.



  * Example1 - "Pen" is a word with one marphem. * Example2 - "Pens" is a word with two marphemes, "pen" and "s".
- **Stem & Root:** The part of a word that an affix is attached to is called a Stem, the word "tie" is root, "untie" is stem.
  - Stemming algorithms work by cutting off the end of the word, and in some cases also the beginning while looking for the root. This indiscriminate cutting can be successful in some occasions, but not always, that is why we affirm that this is an approach that has some limitations.
  - The best-known and most popular stemming approach for English is the Porter stemming algorithm, also known as the Porter stemmer. It is a collection of rules (or, if you prefer, heuristics) designed to reflect how English handles inflections. For example, the Porter stemmer chops both apple and apples down to appl, and it stems berry and berries to berri.
  - Porter Stemmer (5 phase stemming), Lovins Stemmer (1 phase stemming). On English, stemming is not a good idea to convert tokens to terms (recall will be high, precision is going to be low). This technique works well with Spanish, Finish, German, etc., language.

Normalization of terms

- Map single/multiple tokens with terms. Example map window - window, windows; de-accented term - resume.
- Normalization is language dependent
- Case folding - reduce all letters to lowercase
- Synonyms (different words but same meaning) - car = automobile, vehicle; need to map car, automobile, vehicle to automobile.
- Homonyms (same word but different meaning) - saw can be past tence of see or machine cutting wood.
- Spelling mistakes - Soundex is the algorithm, this algorithm will look in to words sound similiar and make them a group.
- Stemming, Lemmatization