aa-14. 自然言語処理

（人工知能）

URL: https://www.kkaneko.jp/ai/mi/index.html

金子邦彦

① **自然言語処理の基礎から最新応用まで。**具体的な例の豊富な提示

② 自然言語処理の基本概念（**構文解析、意味解析**など）。**先進的応用（AI翻訳、画像生成）**。単語の意味的類似性分析（Word2Vec）

③ 獲得できる知識：**自然言語処理の基礎。事前言語処理の原理。Webスクレイピング。意味的類似性**

④ 実践的スキル：Pythonによる自然言語処理プログラミング。Webページからのテキストデータ抽出。自然言語処理モデルの使用

14-1 自然言語処理

自然言語処理

自然言語処理は、**人間が使用する自然言語**を**コンピュータが処理**する技術

① レポートの作成：誤字や脱字を探す

② 文献の下調べ： 翻訳や要約作成

③ 外国語の学習：通訳や音声読み上げ

④ 音声アシスタント：コンピュータへの指示を音声で行う

など

自然言語処理の意義

実際の課題解決に応用可能

例：自動要約システムの開発、長文レポートの効率的な把握

例：大量の顧客フィードバックを感情分析。製品改善に活用

自然言語処理のバリエーション

1. **構文解析**：文の構造を解析し、その部分（例えば、名詞、動詞、形容詞など）を識別
2. **意味解析**：文脈に基づいて単語の意味を理解
3. **翻訳**：ある言語の文を別の言語に変換
4. **品詞の判定**：文中の単語がどの品詞（名詞、動詞、形容詞など）に該当するかを特定
5. **単語数の数え上げ**：文章中の単語数をカウント
6. **音声生成**：テキストを音声に変換
7. **テキスト情報の整理や分類**
8. **自然言語による AI との対話**

WikiPedia の SpaceX の記事

https://en.wikipedia.org/wiki/SpaceX

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

SpaceX

From Wikipedia, the free encyclopedia

Coordinates: 33.9207°N 118.3278°W

This article is about the rocket and spacecraft manufacturer. For the British art gallery, see Spacex (art gallery).
"Space Exploration Technologies" redirects here. For the general topics, see Space exploration and Space technology.

Space Exploration Technologies Corp. (doing business as **SpaceX**) is an American spacecraft manufacturer, space launch provider, and a satellite communications corporation headquartered in Hawthorne, California. SpaceX was founded in 2002 by Elon Musk, with the goal of reducing space transportation costs to enable the colonization of Mars. SpaceX manufactures the Falcon 9 and Falcon Heavy launch vehicles, several rocket engines, Cargo Dragon, crew spacecraft, and Starlink communications satellites.

SpaceX is developing a satellite internet constellation named Starlink to provide commercial internet service. In January 2020, the Starlink constellation became the largest satellite constellation ever launched, and as of May 2022 it comprises over 2,400 small satellites in orbit.[7] The company is also developing Starship, a privately funded, fully reusable, super heavy-lift launch system for interplanetary and orbital spaceflight. Starship is intended to become SpaceX's primary orbital vehicle once operational, supplanting the existing Falcon 9, Falcon Heavy, and Dragon fleet. Starship will have the highest payload capacity of any orbital rocket ever built on its debut, scheduled for 2022 pending launch license.[8]

SpaceX's achievements include the first privately funded liquid-propellant rocket to reach orbit around Earth,[9] the first private company to successfully launch, orbit, and recover a spacecraft, the first private company to send a spacecraft to the International Space Station, the first vertical take-off and vertical propulsive landing for an orbital rocket booster, first reuse of such booster, and the first private company to send astronauts to orbit and to the International Space Station. SpaceX has flown and landed the Falcon 9 series of rockets over one hundred times.

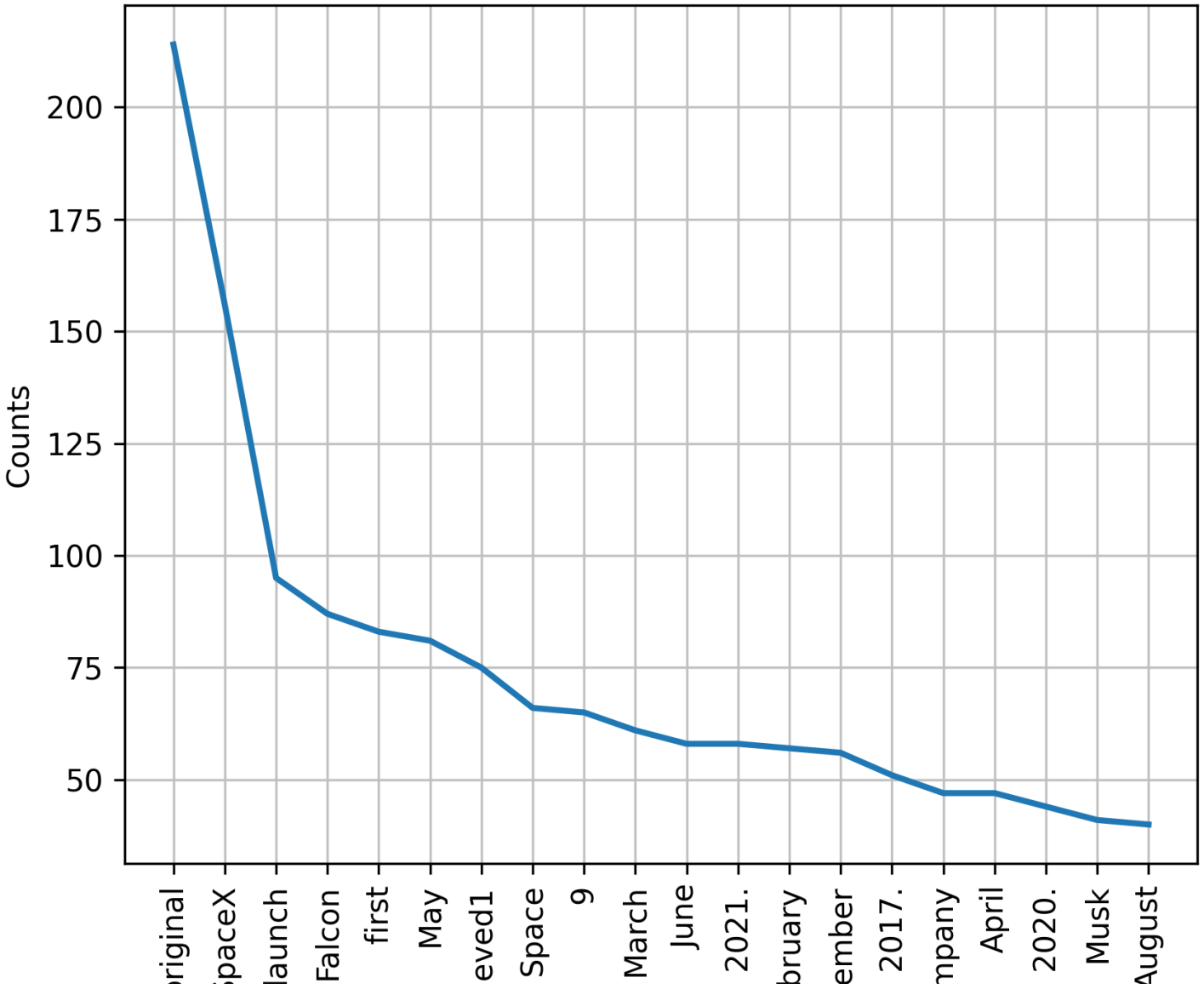
Space Exploration Technologies Corp.

Headquarters in December 2017; plumes from a flight of a Falcon 9 rocket are visible overhead

Trade name	SpaceX
Type	Private
Industry	Space, communications

https://en.wikipedia.org/wiki/SpaceX

Web ページの中の**単語の数を数える**
よく登場する単語は何かのデータ



14-2 構文解析，意味解析

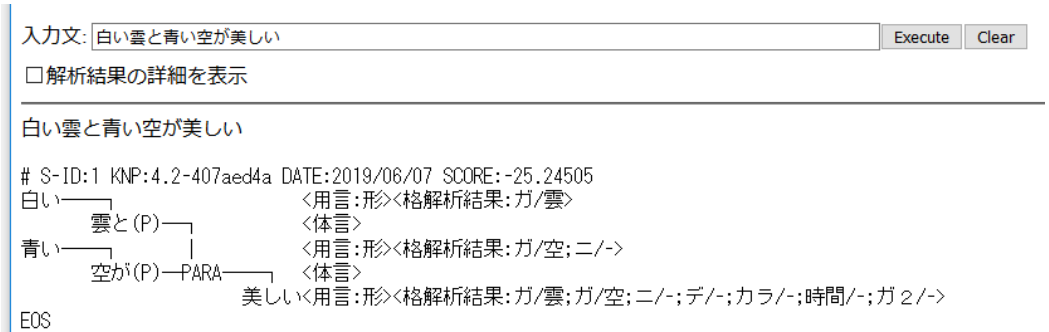
構文解析

構文解析は，自然言語処理の一部で，文章の構造を分析し，文章内の単語の**係り受け**の解析を行う

主語，述語，修飾語などの文章の部分を特定し，それらの間の関係を分析，単語間の関係の分析

- **構文解析** KNP のデモサイト

http://lotus.kuee.kyoto-u.ac.jp/nl-resource/cgi-bin/knp.cgi



演習
構文解析を試してみる

ページ１１、１２

【トピックス】

- **構文解析**
- **係り受け**

演習①

- **構文解析** KNP のデモサイトを開く

http://lotus.kuee.kyoto-u.ac.jp/nl-resource/cgi-bin/knp.cgi

- 自分でいくつかの文章をいれ結果を確認

意味解析の例

- **意味解析**は，自然言語の一部で，コンピュータが文の意味を理解するプロセスである
- 単語やフレーズが，文脈により，どのような意味をもち，どのように解釈されるか分析する

駅前のコンビニは，**油を売っていた**

油：調理油の意味

あの人は，仕事中に，**油を売っていた**

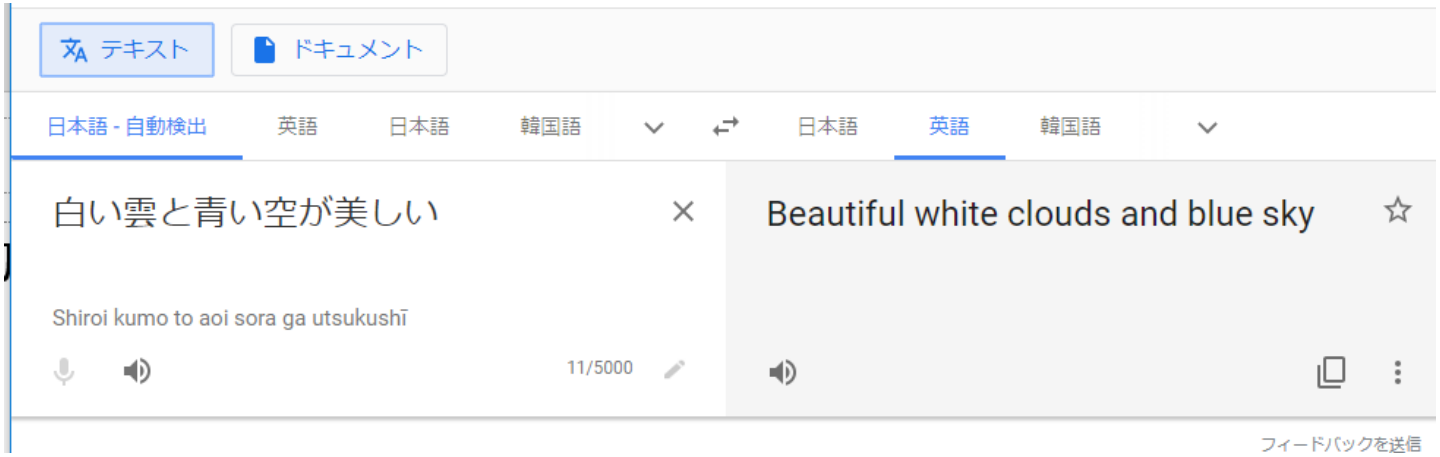
油を売る：仕事をさぼっている

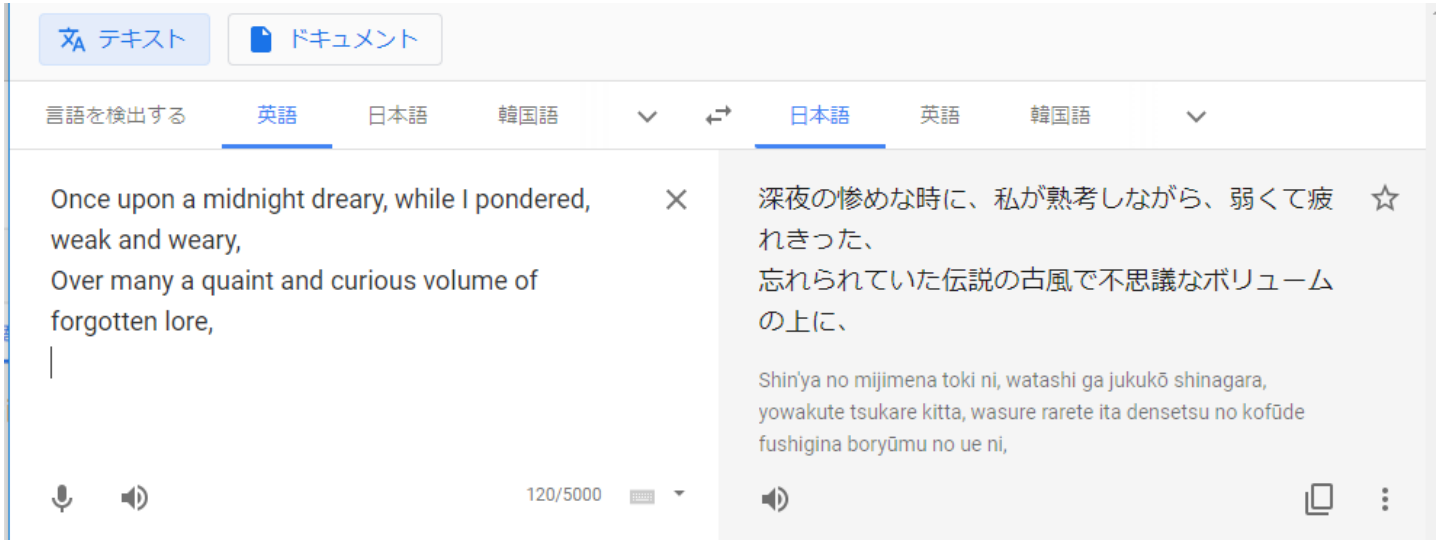
同じ単語やフレーズでも，文脈によって意味が変わる

14-3 人工知能による翻訳

翻訳の例

DeepL https://www.deepl.com//translator





演習
機械翻訳を試してみる

ページ１６、１７

【トピックス】

- **DeepL**

演習②

- DeepL による**翻訳**を試してみる

  DeepL　https://www.deepl.com//translator

- **PDF，ワード，パワーポイントファイル**などの翻訳も可能である

14-4 文章からの画像生成

文章からの画像生成

プロンプトと呼ばれる文章から，短時間で大量の**画像**を生成可能

- プロンプトは，長く詳細に書くなどの工夫が大切
- **倫理的な問題，著作権の尊重**については，利用者で注意が必要

例：生成AIが既存の著作物を学習データとして使用することによる著作権侵害の可能性

例：ディープフェイク技術による人物画像の不正利用

生成AIを画像生成に利用した場合、「生成された画像の出所を明示」するなど、透明性を確保することも重要。

演習
文章からの画像生成

ページ２０，２１

【トピックス】

- **文章からの画像生成**
- **プロンプト**

Demo Stable Diffusion 3 Medium を試す

https://huggingface.co/spaces/stabilityai/stable-diffusion-3-medium

プロンプトを**英語**で入れて「**Run**」をクリック実行

思い通りの結果を得るためにプロンプトを工夫する.

beautiful garden, small house, many flowers, blue sky, clouds, realistic, cinematic, landscape vista photography, Ghibli

Webテキスト処理と分析

Webテキスト処理

- **Webテキスト処理**は、インターネット上の情報を効率的に収集し分析するための役立つ
- 用途の例：市場動向の把握、競合製品の分析、ソーシャルメディアの感情分析など

以下では、Webページからのテキスト抽出と単語頻度分析の基本的な手法を学ぶ。

WikiPedia の SpaceX の記事

https://en.wikipedia.org/wiki/SpaceX

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

SpaceX

From Wikipedia, the free encyclopedia

Coordinates: 33.9207°N 118.3278°W

This article is about the rocket and spacecraft manufacturer. For the British art gallery, see Spacex (art gallery).
"Space Exploration Technologies" redirects here. For the general topics, see Space exploration and Space technology.

Space Exploration Technologies Corp. (doing business as **SpaceX**) is an American spacecraft manufacturer, space launch provider, and a satellite communications corporation headquartered in Hawthorne, California. SpaceX was founded in 2002 by Elon Musk, with the goal of reducing space transportation costs to enable the colonization of Mars. SpaceX manufactures the Falcon 9 and Falcon Heavy launch vehicles, several rocket engines, Cargo Dragon, crew spacecraft, and Starlink communications satellites.

SpaceX is developing a satellite internet constellation named Starlink to provide commercial internet service. In January 2020, the Starlink constellation became the largest satellite constellation ever launched, and as of May 2022 it comprises over 2,400 small satellites in orbit.[7] The company is also developing Starship, a privately funded, fully reusable, super heavy-lift launch system for interplanetary and orbital spaceflight. Starship is intended to become SpaceX's primary orbital vehicle once operational, supplanting the existing Falcon 9, Falcon Heavy, and Dragon fleet. Starship will have the highest payload capacity of any orbital rocket ever built on its debut, scheduled for 2022 pending launch license.[8]

SpaceX's achievements include the first privately funded liquid-propellant rocket to reach orbit around Earth,[9] the first private company to successfully launch, orbit, and recover a spacecraft, the first private company to send a spacecraft to the International Space Station, the first vertical take-off and vertical propulsive landing for an orbital rocket booster, first reuse of such booster, and the first private company to send astronauts to orbit and to the International Space Station. SpaceX has flown and landed the Falcon 9 series of rockets over one hundred times.

Space Exploration Technologies Corp.

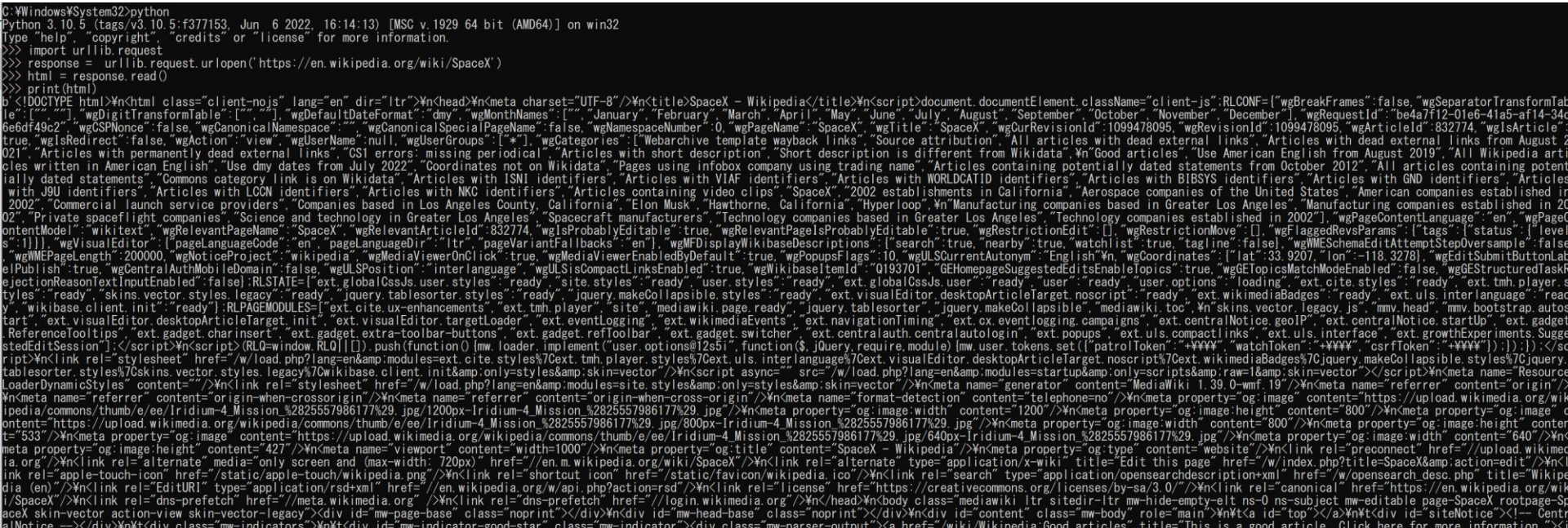
Headquarters in December 2017; plumes from a flight of a Falcon 9 rocket are visible overhead

Trade name	SpaceX
Type	Private
Industry	Space, communications

https://en.wikipedia.org/wiki/SpaceX

Web ページのダウンロードを行うプログラム

```
import urllib.request
response =
urllib.request.urlopen('https://en.wikipedia.org/wiki/SpaceX')
html = response.read()
print(html)
```



この資料のソースコードは，次で公開されているものを使用

https://towardsdatascience.com/gentle-start-to-natural-language-processing-using-python-6e46c07addf3

Web ページから HTML タグを取りのぞくプログラム（テキストが残る）

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html,'html5lib')
text = soup.get_text(strip = True)
print(text)
```

```
>>> from bs4 import BeautifulSoup
>>> soup = BeautifulSoup(html,'html5lib')
>>> text = soup.get_text(strip = True)
>>> print(text)
SpaceX - Wikipediadocument.documentElement.className="client-js";RLCONF={"wgBreakFrames":false,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","April","May","June","July","August","September","October","November","December"],"wgRequestId":"be4a7f12-01e6-41a5-af14-34c6e6df49c2","wgCSPNonce":false,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"SpaceX","wgTitle":"SpaceX","wgCurRevisionId":1099478095,"wgRevisionId":1099478095,"wgArticleId":832774,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["Webarchive template wayback links","Source attribution","All articles with dead external links","Articles with dead external links from August 2021","Articles with permanently dead external links","CS1 errors: missing periodical","Articles with short description","Short description is different from Wikidata",
"Good articles","Use American English from August 2019","All Wikipedia articles written in American English","Use dmy dates from July 2022","Coordinates not on Wikidata","Pages using infobox company using trading name","Articles containing potentially dated statements from October 2012","All articles containing potentially dated statements","Commons category link is on Wikidata","Articles with ISNI identifiers","Articles with VIAF identifiers","Articles with WORLDCATID identifiers","Articles with BIBSYS identifiers","Articles with GND identifiers","Articles with J9U identifiers","Articles with LCCN identifiers","Articles with NKC identifiers","Articles containing video clips","SpaceX","2002 establishments in California","Aerospace companies of the United States","American companies established in 2002","Commercial launch service providers","Companies based in Los Angeles County, California","Elon Musk","Hawthorne, California","Hyperloop",
"Manufacturing companies based in Greater Los Angeles","Manufacturing companies established in 2002","Private spaceflig
```

文章を，単語ごとに切り分けるプログラム

```
tokens = [t for t in text.split()]
print(tokens)
```

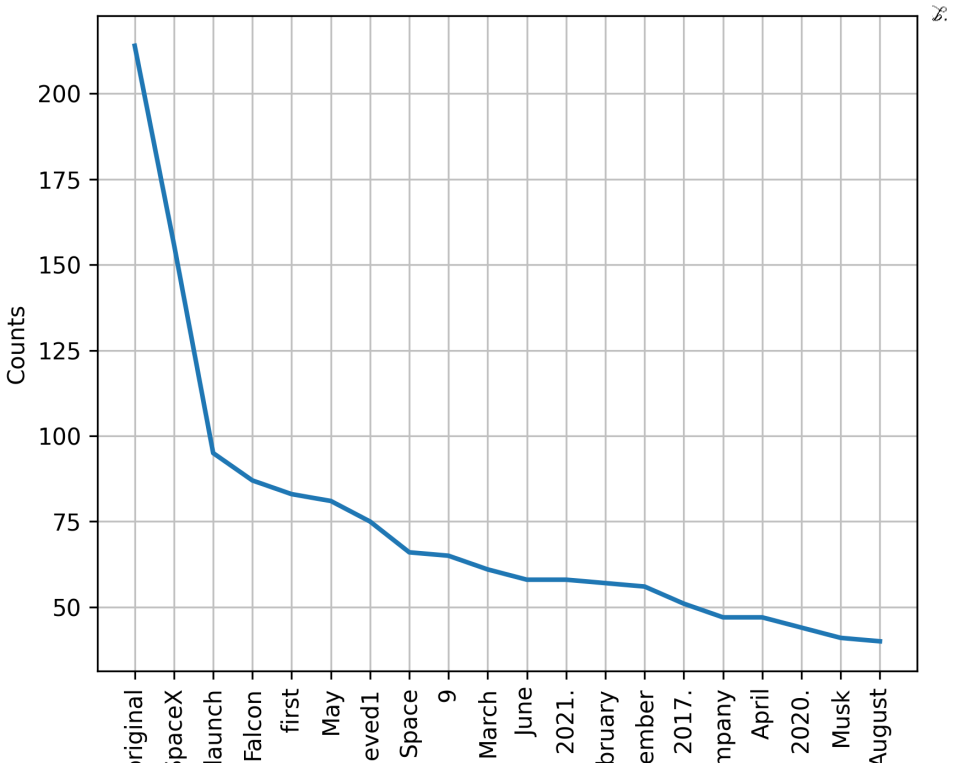
```
>>> tokens = [t for t in text.split()]
>>> print(tokens)
['SpaceX', '-', 'Wikipediadocument.documentElement.className="client-js";RLCONF={"wgBreakFrames":false,"wgSeparatorTrans
formTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","
March","April","May","June","July","August","September","October","November","December"],"wgRequestId":"be4a7f12-01e6-41
a5-af14-34c6e6df49c2","wgCSPNonce":false,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber
":0,"wgPageName":"SpaceX","wgTitle":"SpaceX","wgCurRevisionId":1099478095,"wgRevisionId":1099478095,"wgArticleId":832774
,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["Webar
chive', 'template', 'wayback', 'links","Source', 'attribution","All', 'articles', 'with', 'dead', 'external', 'links","A
rticles', 'with', 'dead', 'external', 'links', 'from', 'August', '2021","Articles', 'with', 'permanently', 'dead', 'exte
rnal', 'links","CS1', 'errors:', 'missing', 'periodical","Articles', 'with', 'short', 'description","Short', 'descriptio
n', 'is', 'different', 'from', 'Wikidata", ', '"Good', 'articles","Use', 'American', 'English', 'from', 'August', '2019",
"All', 'Wikipedia', 'articles', 'written', 'in', 'American', 'English","Use', 'dmy', 'dates', 'from', 'July', '2022","Co
ordinates', 'not', 'on', 'Wikidata","Pages', 'using', 'infobox', 'company', 'using', 'trading', 'name","Articles', 'cont
aining', 'potentially', 'dated', 'statements', 'from', 'October', '2012","All', 'articles', 'containing', 'potentially',
'dated', 'statements","Commons', 'category', 'link', 'is', 'on', 'Wikidata","Articles', 'with', 'ISNI', 'identifiers","
Articles', 'with', 'VIAF', 'identifiers","Articles', 'with', 'WORLDCATID', 'identifiers","Articles', 'with', 'BIBSYS',
identifiers","Articles', 'with', 'GND', 'identifiers","Articles', 'with', 'J9U', 'identifiers","Articles', 'with', 'LCCN
', 'identifiers","Articles', 'with', 'NKC', 'identifiers","Articles', 'containing', 'video', 'clips","SpaceX","2002', 'e
```

単語の数を数えるプログラム

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
sr= stopwords.words('english')
clean_tokens = tokens[:]
for token in tokens:
    if token in stopwords.words('english'):
        clean_tokens.remove(token)

freq = nltk.FreqDist(clean_tokens)
for key,val in freq.items():
    print(str(key) + ':' + str(val))

freq.plot(20, cumulative=False)
```



14-7 単語の意味的類似性分析

単語の分散表現

単語の意味を数値ベクトル（**意味ベクトル**）で表現　＝「**単語の分散表現**」

- 単語間の意味的関係をコンピュータで計算できるようになる。
- コンピュータによる類似語の検出、文脈理解に役立つ

《分散表現の基本的なアイデア》

「似た文脈で使われる単語は、似た意味を持つ」という仮説

例：**「王」と「女王」**、**「男」と「女」**といった単語は、類似した文脈で使用されることが多く、その**意味ベクトルも近くなる**

単語の分散表現の手法 Word2Vec

- Word2Vecは、単語の分散表現を学習する人工知能（AI）
- 近い単語同士が近い位置に配置される

　類似語の発見: **「東京」に近い単語**として「**大阪**」「名古屋」などの他の都市名が得られる

- 単語の関係性の捕捉:

　「王」 - 「男」 + 「女」 = 「女王」といった演算が可能

- 文脈に応じた単語の意味理解:

　「**バンク**」が**金融機関**を指すのか、**川岸**を指すのかを**周辺の単語から判断**

Word2Vec の応用

- 検索エンジンの精度向上
- 推薦システムの改善
- 機械翻訳の品質向上
- 商品推薦
- チャットボット（ChatGPTなど）の応答生成

ここで行うこと

文章の組織化による，人工知能による**類似語の取得**を行う.

- 謝辞

次のサイトの記事，ソースコード，データを使用しています

https://aial.shiroyagi.co.jp/2017/02/japanese-word2vec-model-builder/

```
In [1]: from gensim.models.word2vec import Word2Vec
   ...: mpath = 'word2vec.gensim.model'
   ...: m = Word2Vec.load(mpath)
   ...: def sim(w):
   ...:     for i in m.wv.most_similar(positive=[w]):
   ...:         print(i)
   ...:
   ...: sim('楽しみ')
('楽しく', 0.8377718925476074)
('楽しい', 0.8144588470458984)
('遊び', 0.8099892139434814)
('気軽', 0.797997236251831)
('お客さん', 0.7635385394096375)
('自慢', 0.7526886463165283)
('笑顔', 0.7431896328926086)
('のんびり', 0.7385178804397583)
('大切', 0.7302102446556091)
('お喋り', 0.729663610458374)

In [2]:
```

```
In [3]: sim('晴れ')
('雨', 0.8238000869750977)
('小雨', 0.7920332551002502)
('快晴', 0.7777684926986694)
('降る', 0.7666660547256 47)
('晴天', 0.7512480020523071)
('寒い', 0.7504522800445557)
('晴れる', 0.7483686208724976)
('曇天', 0.7421630024909973)
('曇り', 0.7403557300567627)
('真冬', 0.7319853901863098)

In [4]:
```

Word2Vec により「楽しい」に意味の似ているものを検索

```
In [4]: sim('徳川家康')
('豊臣秀吉', 0.9480844140052795)
('織田信長', 0.929428756237O3)
('毛利輝元', 0.9221905469894409)
('伊達政宗', 0.9190208911895752)
('政宗', 0.9181203842163086)
('上杉景勝', 0.9160507917404175)
('秀吉', 0.9116554260253906)
('徳川氏', 0.9086129069328308)
('今川義元', 0.9033724069595337)
('信玄', 0.8999422788619995)

In [5]:
```

全体まとめ

- **自然言語処理**：人間が使用する言語をコンピュータが処理する技術。誤字や脱字の検出、翻訳や要約作成、音声アシスタントなど、多岐にわたる用途が存在。
- **構文解析**：文の構造を解析し各部分を識別。
- **意味解析**：文脈に基づいて単語の意味を理解。
- 自然言語処理の進歩により、**文章からの画像生成**や**AIによる翻訳**も可能となった
- **テキスト情報の整理や分類**を行うことも可能

①**自然言語処理の学習意義と価値**。自然言語処理は多数の応用。デジタル社会の基盤技術。**最新AIサービス（DeepL翻訳、ChatGPTなど）の基礎**

② 技術的スキルの向上：実際に動作させることによる自然言語処理活用スキルの向上、**学習による着実な成長と達成感**

③ AI技術の可能性と創造的応用：**AIと人間の対話によるコミュニケーション、文書の推敲、文章からの画像生成など**、AI技術の革新的応用。

④ 未来への展望：ビッグデータ分析やAI開発分野での可能性、総合的能力の獲得