

Web Scraping a Wiki for Summarization

By: Eric Tran

Introduction

In this project, I will use various Python libraries to webscrape, or acquire the text content of a wikipedia page. I will then use various Python libraries and NLP techniques to clean the data (EDA), perform named entity recognition (NER), part of speech tagging (POS), tokenization to prepare the data, and wordcloud to visualize the top words in the document. I will then use transformers to summarize and perform sentiment analysis.

I will then attempt to make a simple chatbot, with the added feature that it will get a sentiment analysis on the user input as well as the chatbot response. It will then create a WordCloud of the top 10 words of the chatbot response, as well as return a summary of the wikipedia article for the main subject of the chatbot response.

First we will need to find a wikipedia page. We already know that Wikipedia pages start with <https://en.wikipedia.org/wiki/> (<https://en.wikipedia.org/wiki/>), and so we will save that to a variable and append a topic. We will start by importing our libraries.

```
In [35]: # processing Libraries
import nltk
from nltk.corpus import stopwords
nltk.download("punkt")
nltk.download("stopwords")
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.probability import FreqDist
from textblob import TextBlob # for sentiment analysis
import numpy as np
import regex # for regular expressions
import requests # for getting HTML code
from bs4 import BeautifulSoup # for parsing the HTML code

# neural network and model Libraries
import tensorflow as tf # for tensor use
from tensorflow.keras.layers import Input, LSTM, Dense
from tensorflow.keras.models import Model

# visualization Libraries
import matplotlib.pyplot as plt
from wordcloud import WordCloud
import spacy
from spacy import displacy
```

WebScraping

The first thing we are going to do is scrape a wikipedia page. We will use requests to scrape the page for html code, then parse and save the response in a BeautifulSoup object to extract the paragraphs. We will then concatenate the paragraphs to create a long string with all the paragraph texts.

```
In [15]: no_wiki_msg = "Sorry, there is no wikipedia page for your subject." # this message will display if there is no wikipedia entr
valid_url = False # this is trigger for whether search string is valid or not
wiki_path = 'https://en.wikipedia.org/wiki/' # general path for wikipedia articles
while not valid_url:
    user_input = str(input('Please Enter a Wikipedia subject: ')) # collects user input
    wiki_url = wiki_path + user_input # create url using input
    response = requests.get(wiki_url) # retrieve url response
    html_content = response.content # retrieve content from response
    soup = BeautifulSoup(html_content, 'html.parser') # create a beautiful soup object with the content
    paragraphs = soup.find_all('p') # find all paragraphs in content
    paragraph_text = [paragraph.get_text() for paragraph in paragraphs] # get text for each paragraph
    all_text = ''
    for text in paragraph_text: # iterates thru all the paragraphs
        all_text = all_text + text # adds each paragraph as text creating giant string
    if all_text == 'Other reasons this message may be displayed:\n': # checks if wiki page is not valid
        print(no_wiki_msg) # if not valid, returns error message and loops starts over
    else:
        print(f'{user_input} is a valid wikipedia subject!') # if valid, prints valid message and set trigger to true, escap
        valid_url = True
```

```
Please Enter a Wikipedia subject: gold
gold is a valid wikipedia subject!
```

In [16]:  `print(all_text)`

Gold is a chemical element; it has symbol Au (from Latin aurum 'gold') and atomic number 79. In pure form, it is a bright, slightly orange-yellow, dense, soft, malleable, and ductile metal. Chemically, gold is a transition metal, a group 11 element, and one of the noble metals. It is one of the least reactive chemical elements and is solid under standard conditions. Gold often occurs in free elemental (native state), as nuggets or grains, in rocks, veins, and alluvial deposits. It occurs in a solid solution series with the native element silver (as in electrum), naturally alloyed with other metals like copper and palladium, and mineral inclusions such as within pyrite. Less commonly, it occurs in minerals as gold compounds, often with tellurium (gold tellurides).

Gold is resistant to most acids, though it does dissolve in aqua regia (a mixture of nitric acid and hydrochloric acid), forming a soluble tetrachloroaurate anion. Gold is insoluble in nitric acid alone, which dissolves silver and base metals, a property long used to refine gold and confirm the presence of gold in metallic substances, giving rise to the term 'acid test'. Gold dissolves in alkaline solutions of cyanide, which are used in mining and electroplating. Gold also dissolves in mercury, forming amalgam alloys, and as the gold acts simply as a solute, this is not a chemical reaction.

A relatively rare element,[7][8] gold is a precious metal that has been used for coinage, jewelry, and other arts throughout recorded history. In the past, a gold standard was often implemented as a monetary policy. Gold coins ceased to be minted as a circulating currency in the 1930s, and the world gold standard was abandoned for a fiat currency system after the Nixon shock measures of 1971.

In 2020, the world's largest gold producer was China, followed by Russia and Australia.[9] A total of around 201,296 tonnes of gold exists above ground, as of 2020[update].[10] This is equal to a cube with each side measuring roughly 21.7 meters (71 ft). The world's consumption of new gold produced is about 50% in jewelry, 40% in investments, and 10% in industry.[11] Gold's high malleability, ductility, resistance to corrosion and most other chemical reactions, and conductivity of electricity have led to its continued use in corrosion-resistant electrical connectors in all types of computerized devices (its chief industrial use). Gold is also used in infrared shielding, the production of colored glass, gold leafing, and tooth restoration. Certain gold salts are still used as anti-inflammatories in medicine.

Gold is the most malleable of all metals. It can be drawn into a wire of single-atom width, and then stretched considerably before it breaks.[12] Such nanowires distort via the formation, reorientation, and migration of dislocations and crystal twins without noticeable hardening.[13] A single gram of gold can be beaten into a sheet of 1 square metre (11 sq ft), and an avoirdupois ounce into 300 square feet (28 m²). Gold leaf can be beaten thin enough to become semi-transparent. The transmitted light appears greenish-blue because gold strongly reflects yellow and red.[14] Such semi-transparent sheets also strongly reflect infrared light, making them useful as infrared (radiant heat) shields in the visors of heat-resistant suits and in sun visors for spacesuits.[15] Gold is a good conductor of heat and electricity.

Gold has a density of 19.3 g/cm³, almost identical to that of tungsten at 19.25 g/cm³; as such, tungsten has been used in the counterfeiting of gold bars, such as by plating a tungsten bar with gold.[16][17][18][19] By comparison, the density of lead is 11.34 g/cm³, and that of the densest element, osmium, is 22.588±0.015 g/cm³. [20]

Whereas most metals are gray or silvery white, gold is slightly reddish-yellow.[21] This color is determined by the frequency of plasma oscillations among the metal's valence electrons, in the ultraviolet range for most metals but in the visible range for gold due to relativistic effects affecting the orbitals around gold atoms.[22][23] Similar effects impart a golden hue to metallic caesium.

Common colored gold alloys include the distinctive eighteen-karat rose gold created by the addition of copper. Alloys containing palladium or nickel are also important in commercial jewelry as these produce white gold alloys. Fourteen-karat gold-copper alloy is nearly identical in color to certain bronze alloys, and both may be used to produce police and other badges. Fourteen- and eighteen-karat gold alloys with silver alone appear greenish-yellow and are referred to as green gold. Blue gold can be made by alloying with iron, and purple gold can be made by alloying with aluminium. Less commonly, addition of manganese, indium, and other elements can produce more unusual colors of gold for various applications.[24] Colloidal gold, used by electron-microscopists, is red if the particles are small; larger particles of colloidal gold are blue.[25]

Gold has only one stable isotope, ¹⁹⁷Au, which is also its only naturally occurring isotope, so gold is both a mononuclidic and monoisotopic element. Thirty-six radioisotopes have been synthesized, ranging in atomic mass from 169 to 205. The most stable of these is ¹⁹⁵Au with a half-life of 186.1 days. The least stable is ¹⁷¹Au, which decays by proton emission with a half-life of 30 μs. Most of gold's radioisotopes with atomic masses below 197 decay by some combination of proton emission, α decay, and β⁺ decay. The exceptions are ¹⁹⁵Au, which decays by electron capture, and ¹⁹⁶Au, which decays most often by electron capture (93%) with a minor β[−] decay path (7%).[26] All of gold's radioisotopes with atomic masses above 197 decay by β[−] decay.[27]

At least 32 nuclear isomers have also been characterized, ranging in atomic mass from 170 to 200. Within that range, only ¹⁷⁸Au, ¹⁸⁰Au, ¹⁸¹Au, ¹⁸²Au, and ¹⁸⁸Au do not have isomers. Gold's most stable isomer is ^{198m2}Au with a half-life of 2.27 days. Gold's least stable isomer is ^{177m2}Au with a half-life of only 7 ns. ^{184m1}Au has three decay paths: β⁺ decay, isomeric transition, and alpha decay. No other isomer or isotope of gold has three decay paths.[27]

The possible production of gold from a more common element, such as lead, has long been a subject of human inquiry, and the ancient and medieval discipline of alchemy often focused on it; however, the transmutation of the chemical elements did not become possible until the understanding of nuclear physics in the 20th century. The first synthesis of gold was conducted by Japanese physicist Hantaro Nagaoka, who synthesized gold from mercury in 1924 by neutron bombardment.[28] An American team, working without knowledge of Nagaoka's prior study, conducted the same experiment in 1941, achieving the same result and showing that the isotopes of gold produced by it were all radioactive.[29] In 1980, Glenn Seaborg transmuted several thousand atoms of bismuth into gold at the Lawrence Berkeley Laboratory.[30][31] Gold can be manufactured in a nuclear reactor, but doing so is highly impractical and would cost far more than the value of the gold that is produced.[32]

Although gold is the most noble of the noble metals,[33][34] it still forms many diverse compounds. The oxidation state of gold in its compounds ranges from −1 to +5, but Au(I) and Au(III) dominate its chemistry. Au(I), referred to as the aurous ion, is the most common oxidation state with soft ligands such as thioethers, thiolates, and organophosphines. Au(I) compounds are typically linear. A good example is Au(CN)₂[−], which is the soluble form of gold encountered in mining. The binary gold halides, such as AuCl₃, form zigzag polymeric chains, again featuring linear coordination at Au. Most drugs based on gold are Au(I) derivatives.[35]

Au(III) (referred to as auric) is a common oxidation state, and is illustrated by gold(III) chloride, Au₂Cl₆. The gold atom centers in Au(III) complexes, like other d⁸ compounds, are typically square planar, with chemical bonds that have both covalent and ionic character. Gold(I,III) chloride is also known, an example of a mixed-valence complex.

Gold does not react with oxygen at any temperature[36] and, up to 100 °C, is resistant to attack from ozone:[37]

Au
+

0

2

```

→
(
no reaction
)

{\displaystyle {\ce {Au + O2 ->}}{\text{no reaction}}}}

```

Au

+

O

3

→

t

<

100

°

C

```

(
no reaction
)

{\displaystyle {\ce {Au}+O3->[\text{atop }t<100^{\circ }]{\text{C}}}}{\text{no reaction}}}

```

Some free halogens react with gold.[38] Gold is strongly attacked by fluorine at dull-red heat[39] to form gold(III) fluoride AuF₃. Powdered gold reacts with chlorine at 180 °C to form gold(III) chloride AuCl₃. [40] Gold reacts with bromine at 140 °C to form gold(III) bromide AuBr₃, but reacts only very slowly with iodine to form gold(I) iodide AuI:

2

Au

+

3

F

2

→

Δ

2

AuF

3

$$\{ \backslash displaystyle \{ \backslash ce {2Au\{ }+3F2-\>[\{ \} \ atop \Delta]2AuF3\} \}$$

2

Au

+

3

Cl

2

→

Δ

2

AuCl

3

```
{\displaystyle {\ce {2Au{}}+3Cl2->[{} \atop \Delta ]2AuCl3}}}
```

2

Au

+

2

Br

2

→

Δ

AuBr

3

+

AuBr

```
{\displaystyle {\ce {2Au{}}+2Br2->[{} \atop \Delta ]AuBr3{}}+AuBr{}}}
```

2

Au

+

I

2

→

Δ

2

AuI

$${\ce {2Au} + I2 ->[{} \atop \Delta] 2AuI}}$$

Gold does not react with sulfur directly,[41] but gold(III) sulfide can be made by passing hydrogen sulfide through a dilute solution of gold(III) chloride or chlorauric acid.

Unlike sulfur, phosphorus reacts directly with gold at elevated temperatures to produce gold phosphide (Au₂P₃).[42]

Gold readily dissolves in mercury at room temperature to form an amalgam, and forms alloys with many other metals at higher temperatures. These alloys can be produced to modify the hardness and other metallurgical properties, to control melting point or to create exotic colors.[24]

Gold is unaffected by most acids. It does not react with hydrofluoric, hydrochloric, hydrobromic, hydriodic, sulfuric, or nitric acid. It does react with selenic acid, and is dissolved by aqua regia, a 1:3 mixture of nitric acid and hydrochloric acid. Nitric acid oxidizes the metal to +3 ions, but only in minute amounts, typically undetectable in the pure acid because of the chemical equilibrium of the reaction. However, the ions are removed from the equilibrium by hydrochloric acid, forming AuCl₄⁻ ions, or chlorauric acid, thereby enabling further oxidation:

2

Au

+

6

H

2

SeO

4

→

200

°

C

Au

2

(
 SeO
 4

)
 3

 +
 3

 H
 2

 SeO
 3

 +
 3

 H
 2

 0

$$\{ \text{\ce{2Au}} + 6\text{H2SeO4} \rightarrow [{} \text{\atop {200^{\circ}} \text{\text{C}}}] \text{Au2(SeO4)3} + 3\text{H2SeO3} + 3\text{H2O} \}$$

 Au

 +
 4
 HCl

 +
 HNO
 3

→

HAuCl

4

+

NO

↑

+

2

H

2

0

$$\{\text{Au}\} + 4\text{HCl} + \text{HNO}_3 \rightarrow \text{HAuCl}_4 + \text{NO} \uparrow + 2\text{H}_2\text{O}$$

Gold is similarly unaffected by most bases. It does not react with aqueous, solid, or molten sodium or potassium hydroxide. It does however, react with sodium or potassium cyanide under alkaline conditions when oxygen is present to form soluble complexes.[41]

Common oxidation states of gold include +1 (gold(I) or aurous compounds) and +3 (gold(III) or auric compounds). Gold ions in solution are readily reduced and precipitated as metal by adding any other metal as the reducing agent. The added metal is oxidized and dissolves, allowing the gold to be displaced from solution and be recovered as a solid precipitate.

Less common oxidation states of gold include -1, +2, and +5.

The -1 oxidation state occurs in aurides, compounds containing the Au⁻ anion. Caesium auride (CsAu), for example, crystallizes in the caesium chloride motif;[43] rubidium, potassium, and tetramethylammonium aurides are also known.[44] Gold has the highest electron affinity of any metal, at 222.8 kJ/mol, making Au⁻ a stable species,[45] analogous to the halides.

Gold also has a -1 oxidation state in covalent complexes with the group 4 transition metals, such as in titanium tetraauride and the analogous zirconium and hafnium compounds. These chemicals are expected to form gold-bridged dimers in a manner similar to titanium(IV) hydride.[46]

Gold(II) compounds are usually diamagnetic with Au-Au bonds such as [Au(CH₂)₂P(C₆H₅)₂]₂Cl₂. The evaporation of a solution of Au(OH)₃ in concentrated H₂SO₄ produces red crystals of gold(II) sulfate, Au₂(SO₄)₂. Originally thought to be a mixed-valence compound, it has been shown to contain Au⁴⁺ cations, analogous to the better-known mercury(I) ion, Hg₂²⁺. [47][48] A gold(I) complex, the tetraxenonogold(II) cation, which contains xenon as a ligand, occurs in [AuXe₄](Sb₂F₁₁)₂. [49] In September 2023, a novel type of metal-halide perovskite material consisting of Au³⁺ and Au²⁺ cations in its crystal structure has been found.[50] It has been shown to unexpectedly stable at normal conditions.

Gold pentafluoride, along with its derivative anion, AuF₆⁻, and its difluorine complex, gold heptafluoride, is the sole example of gold(V), the highest verified oxidation state.[51]

Some gold compounds exhibit aurophilic bonding, which describes the tendency of gold ions to interact at distances that are too long to be a conventional Au-Au bond but shorter than van der Waals bonding. The interaction is estimated to be comparable in strength to that of a hydrogen bond.

Well-defined cluster compounds are numerous.[44] In some cases, gold has a fractional oxidation state. A representative example is the octahedral species {Au(P(C₆H₅)₃)₂}₆.

Gold is thought to have been produced in supernova nucleosynthesis, and from the collision of neutron stars,[52] and to have been present in the dust from which the Solar System formed.[53]

Traditionally, gold in the universe is thought to have formed by the r-process (rapid neutron capture) in supernova nucleosynthesis,[54] but more recently it has been suggested that gold and other elements heavier than iron may also be produced in quantity by the r-process in the collision of neutron stars.[55] In both cases, satellite spectrometers at first only indirectly detected the resulting gold.[56] However, in August 2017, the spectroscopic signatures of heavy elements, including gold, were observed by electromagnetic observatories in the GW170817 neutron star merger event, after gravitational wave detectors confirmed the event as a neutron star merger.[57] Current astrophysical models suggest that this single neutron star merger event generated between 3 and 13 Earth masses of gold. This amount, along with estimations of the rate of occurrence of these neutron star merger events, suggests that such mergers may produce enough gold to account for most of the abundance of this element in the universe.[58]

Because the Earth was molten when it was formed, almost all of the gold present in the early Earth probably sank into the planetary core. Therefore, most of the gold that is in the Earth's crust and mantle has in one model thought to have been delivered to Earth later, by asteroid impacts during the Late Heavy Bombardment, about 4 billion years ago.[59][60]

Gold which is reachable by humans has, in one case, been associated with a particular asteroid impact. The asteroid that formed Vredefort impact structure 2.020 billion years ago is often credited with seeding the Witwatersrand basin in South Africa with the richest gold deposits on earth.[61][62][63][64] However, this scenario is now questioned. The gold-bearing Witwatersrand rocks were laid down between 700 and 950 million years before the Vredefort impact.[65][66] These gold-bearing rocks had furthermore been covered by a thick layer of Ventersdorp lavas and the Transvaal Supergroup of rocks before the meteor strike, and thus the gold did not actually arrive in the asteroid/meteorite. What the Vredefort impact achieved, however, was to distort the Witwatersrand basin in such a way that the gold-bearing rocks were brought to the present erosion surface in Johannesburg, on the Witwatersrand, just inside the rim of the original 300 km (190 mi) diameter crater caused by the meteor strike. The discovery of the deposit in 1886 launched the Witwatersrand Gold Rush. Some 22% of all the gold that is ascertained to exist today on Earth has been extracted from these Witwatersrand rocks.[66]

Notwithstanding the impact above, much of the rest of the gold on Earth is thought to have been incorporated into the planet

since its very beginning, as planetesimals formed the planet's mantle, early in Earth's creation. In 2017, an international group of scientists, established that gold "came to the Earth's surface from the deepest regions of our planet", [67] the mantle, evidenced by their findings at Deseado Massif in the Argentinian Patagonia. [68] [clarification needed] On Earth, gold is found in ores in rock formed from the Precambrian time onward. [69] It most often occurs as a native metal, typically in a metal solid solution with silver (i.e. as a gold/silver alloy). Such alloys usually have a silver content of 8–10%. Electrum is elemental gold with more than 20% silver, and is commonly known as white gold. Electrum's color runs from golden-silvery to silvery, dependent upon the silver content. The more silver, the lower the specific gravity.

Native gold occurs as very small to microscopic particles embedded in rock, often together with quartz or sulfide minerals such as "fool's gold", which is a pyrite. [70] These are called lode deposits. The metal in a native state is also found in the form of free flakes, grains or larger nuggets [69] that have been eroded from rocks and end up in alluvial deposits called placer deposits. Such free gold is always richer at the exposed surface of gold-bearing veins, owing to the oxidation of accompanying minerals followed by weathering; and by washing of the dust into streams and rivers, where it collects and can be welded by water action to form nuggets.

Gold sometimes occurs combined with tellurium as the minerals calaverite, krennerite, nagyagite, petzite and sylvanite (see telluride minerals), and as the rare bismuthide maldonite (Au₂Bi) and antimonide aurostibite (AuSb₂). Gold also occurs in rare alloys with copper, lead, and mercury: the minerals auricupride (Cu₃Au), novodneprite (AuPb₃) and weishanite ((Au,Ag)₃Hg₂).

Recent research suggests that microbes can sometimes play an important role in forming gold deposits, transporting and precipitating gold to form grains and nuggets that collect in alluvial deposits. [71]

Another recent study has claimed water in faults vaporizes during an earthquake, depositing gold. When an earthquake strikes, it moves along a fault. Water often lubricates faults, filling in fractures and jogs. About 10 kilometres (6.2 mi) below the surface, under very high temperatures and pressures, the water carries high concentrations of carbon dioxide, silica, and gold. During an earthquake, the fault jog suddenly opens wider. The water inside the void instantly vaporizes, flashing to steam and forcing silica, which forms the mineral quartz, and gold out of the fluids and onto nearby surfaces. [72]

The world's oceans contain gold. Measured concentrations of gold in the Atlantic and Northeast Pacific are 50–150 femtomol/L or 10–30 parts per quadrillion (about 10–30 g/km³). In general, gold concentrations for south Atlantic and central Pacific samples are the same (~50 femtomol/L) but less certain. Mediterranean deep waters contain slightly higher concentrations of gold (100–150 femtomol/L) attributed to wind-blown dust or rivers. At 10 parts per quadrillion the Earth's oceans would hold 15,000 tonnes of gold. [73] These figures are three orders of magnitude less than reported in the literature prior to 1988, indicating contamination problems with the earlier data.

A number of people have claimed to be able to economically recover gold from sea water, but they were either mistaken or acted in an intentional deception. Prescott Jernegan ran a gold-from-seawater swindle in the United States in the 1890s, as did an English fraudster in the early 1900s. [74] Fritz Haber did research on the extraction of gold from sea water in an effort to help pay Germany's reparations following World War I. [75] Based on the published values of 2 to 64 ppb of gold in seawater a commercially successful extraction seemed possible. After analysis of 4,000 water samples yielding an average of 0.004 ppb it became clear that extraction would not be possible and he ended the project. [76]

The earliest recorded metal employed by humans appears to be gold, which can be found free or "native". Small amounts of natural gold have been found in Spanish caves used during the late Paleolithic period, c. 40,000 BC. [78]

The oldest gold artifacts in the world are from Bulgaria and are dating back to the 5th millennium BC (4,600 BC to 4,200 BC), such as those found in the Varna Necropolis near Lake Varna and the Black Sea coast, thought to be the earliest "well-dated" finding of gold artifacts in history. [79] [69] [80] Several prehistoric Bulgarian finds are considered no less old – the golden treasures of Hotnitsa, Durankulak, artifacts from the Kurgan settlement of Yunatsite near Pazardzhik, the golden treasure of Sakar, as well as beads and gold jewelry found in the Kurgan settlement of Provadia – Solnitsata ("salt pit"). However, Varna gold is most often called the oldest since this treasure is the largest and most diverse. [81]

Gold artifacts probably made their first appearance in Ancient Egypt at the very beginning of the pre-dynastic period, at the end of the fifth millennium BC and the start of the fourth, and smelting was developed during the course of the 4th millennium; gold artifacts appear in the archeology of Lower Mesopotamia during the early 4th millennium. [82] As of 1990, gold artifacts found at the Wadi Qana cave cemetery of the 4th millennium BC in West Bank were the earliest from the Levant. [83] Gold artifacts such as the golden hats and the Nebra disk appeared in Central Europe from the 2nd millennium BC Bronze Age.

The oldest known map of a gold mine was drawn in the 19th Dynasty of Ancient Egypt (1320–1200 BC), whereas the first written reference to gold was recorded in the 12th Dynasty around 1900 BC. [84] Egyptian hieroglyphs from as early as 2600 BC describe gold, which King Tushratta of the Mitanni claimed was "more plentiful than dirt" in Egypt. [85] Egypt and especially Nubia had the resources to make them major gold-producing areas for much of history. One of the earliest known maps, known as the Turin Papyrus Map, shows the plan of a gold mine in Nubia together with indications of the local geology. The primitive working methods are described by both Strabo and Diodorus Siculus, and included fire-setting. Large mines were also present across the Red Sea in what is now Saudi Arabia.

Gold is mentioned in the Amarna letters numbered 19 [86] and 26 [87] from around the 14th century BC. [88] [89]

Gold is mentioned frequently in the Old Testament, starting with Genesis 2:11 (at Havilah), the story of the golden calf, and many parts of the temple including the Menorah and the golden altar. In the New Testament, it is included with the gifts of the magi in the first chapters of Matthew. The Book of Revelation 21:21 describes the city of New Jerusalem as having streets "made of pure gold, clear as crystal". Exploitation of gold in the south-east corner of the Black Sea is said to date from the time of Midas, and this gold was important in the establishment of what is probably the world's earliest coinage in Lydia around 610 BC. [90] The legend of the golden fleece dating from eighth century BCE may refer to the use of fleeces to trap gold dust from placer deposits in the ancient world. From the 6th or 5th century BC, the Chu (state) circulated the Ying Yuan, one kind of square gold coin.

In Roman metallurgy, new methods for extracting gold on a large scale were developed by introducing hydraulic mining methods, especially in Hispania from 25 BC onwards and in Dacia from 106 AD onwards. One of their largest mines was at Las Medulas in León, where seven long aqueducts enabled them to sluice most of a large alluvial deposit. The mines at Roşia Montană in Transylvania were also very large, and until very recently, still mined by opencast methods. They also exploited smaller deposits in Britain, such as placer and hard-rock deposits at Dolaucothi. The various methods they used are well described by Pliny the Elder in his encyclopedia *Naturalis Historia* written towards the end of the first century AD.

During Mansa Musa's (ruler of the Mali Empire from 1312 to 1337) hajj to Mecca in 1324, he passed through Cairo in July 1324, and was reportedly accompanied by a camel train that included thousands of people and nearly a hundred camels where he gave away so much gold that it depressed the price in Egypt for over a decade, causing high inflation. [91] A contemporary Arab historian remarked:

Gold was at a high price in Egypt until they came in that year. The mithqal did not go below 25 dirhams and was generally above, but from that time its value fell and it cheapened in price and has remained cheap till now. The mithqal does not exceed 22 dirhams or less. This has been the state of affairs for about twelve years until this day by reason of the large amount of gold which they brought into Egypt and spent there [...]. The European exploration of the Americas was fueled in no small part by reports of the gold ornaments displayed in great profusion by Native American peoples, especially in Mesoamerica, Peru, Ecuador and Colombia. The Aztecs regarded gold as the product of the gods, calling it literally "god excrement" (teocuitlatl in Nahuatl), and after Moctezuma II was killed, most of this gold was shipped to Spain. [94] However, for the indigenous peoples of North America gold was considered useless and they saw much greater value in other minerals which were directly related to their utility, such as obsidian, flint, and slate. [95] El Dorado is applied to a legendary story in which precious stones were found in fabulous abundance along with gold coins. The concept of El Dorado underwent several transformations, and eventually accounts of the previous myth were also combined with those of a legendary lost city. El Dorado, was the term

used by the Spanish Empire to describe a mythical tribal chief (zipa) of the Muisca native people in Colombia, who, as an initiation rite, covered himself with gold dust and submerged in Lake Guatavita. The legends surrounding El Dorado changed over time, as it went from being a man, to a city, to a kingdom, and then finally to an empire.

Beginning in the early modern period, European exploration and colonization of West Africa was driven in large part by reports of gold deposits in the region, which was eventually referred to by Europeans as the "Gold Coast".[96] From the late 15th to early 19th centuries, European trade in the region was primarily focused in gold, along with ivory and slaves.[97] The gold trade in West Africa was dominated by the Ashanti Empire, who initially traded with the Portuguese before branching out and trading with British, French, Spanish and Danish merchants.[98] British desires to secure control of West African gold deposits played a role in the Anglo-Ashanti wars of the late 19th century, which saw the Ashanti Empire annexed by Britain.[99]

Gold played a role in western culture, as a cause for desire and of corruption, as told in children's fables such as Rumpelstiltskin—where Rumpelstiltskin turns hay into gold for the peasant's daughter in return for her child when she becomes a princess—and the stealing of the hen that lays golden eggs in Jack and the Beanstalk.

The top prize at the Olympic Games and many other sports competitions is the gold medal.

75% of the presently accounted for gold has been extracted since 1910, two-thirds since 1950.

One main goal of the alchemists was to produce gold from other substances, such as lead – presumably by the interaction with a mythical substance called the philosopher's stone. Trying to produce gold led the alchemists to systematically find out what can be done with substances, and this laid the foundation for today's chemistry, which can produce gold (albeit uneconomically) by using nuclear transmutation.[100] Their symbol for gold was the circle with a point at its center (⦿), which was also the astrological symbol and the ancient Chinese character for the Sun.

The Dome of the Rock is covered with an ultra-thin golden glassier. The Sikh Golden temple, the Harmandir Sahib, is a building covered with gold. Similarly the Wat Phra Kaew emerald Buddhist temple (wat) in Thailand has ornamental gold-leafed statues and roofs. Some European king and queen's crowns were made of gold, and gold was used for the bridal crown since antiquity. An ancient Talmudic text circa 100 AD describes Rachel, wife of Rabbi Akiva, receiving a "Jerusalem of Gold" (diadem). A Greek burial crown made of gold was found in a grave circa 370 BC.

Minoan jewellery, 2300–2100 BC, gold, Metropolitan Museum of Art, New York

Sumerian earrings with cuneiform inscriptions, 2093–2046 BC, gold, Sulaymaniyah Museum, Sulaymaniyah, Iraq

Minoan cup, part of the Aegina Treasure, 1850–1550 BC, gold, British Museum[101]

Ancient Egyptian statuette of Amun, 945–715 BC, gold, Metropolitan Museum of Art

Ancient Egyptian signet ring, 664–525 BC, gold, British Museum

Ancient Chinese cast openwork dagger hilt, 6th–5th centuries BC, gold, British Museum[102]

Ancient Greek stater, 323–315 BC, gold, Metropolitan Museum of Art

Etruscan funerary wreath, 4th–3rd century BC, gold, Metropolitan Museum of Art

Roman aureus of Hadrian, 134–138 AD, gold, Metropolitan Museum of Art

Quimbaya lime container, 5th–9th century, gold, Metropolitan Museum of Art

Anglo-Saxon belt buckle from Sutton Hoo with a niello interlace pattern, 7th century, gold, British Museum[103]

Byzantine scyphate, 1059–1067, gold, Cleveland Museum of Art, Cleveland, Ohio, USA

Pre-Columbian pendant with two bat-head warriors who carry spears, 11th–16th century, gold, Metropolitan Museum of Art

Inca hollow model of a llama, 14th–15th centuries, gold, British Museum[104]

Renaissance hat badge that shows the Judgment of Paris, 16th century, enamelled gold, British Museum[105]

Rococo box, by George Michael Moser, 1741, gold, Metropolitan Museum of Art

Rococo candelabrum, by Jean Joseph de Saint-Germain, c.1750, gilt bronze, Cleveland Museum of Art

Rococo snuff box with Minerva, by Jean-Malquis Lequin, 1750–1752, gold and painted enamel, Louvre[106]

Louis XVI style snuff box, by Jean Frémin, 1763–1764, gold and painted enamel, Louvre[107]

Neoclassical washstand (athénienne or lavabo), 1800–1814, legs, base and shelf of yew wood, gilt bronze mounts, iron plate beneath shelf, Metropolitan Museum of Art

Gothic Revival clock, unknown French maker, c.1835–1840, gilt and patinated bronze, Museum of Decorative Arts, Paris

Art Nouveau teapot, by Alphonse Debain, gilt silver and ivory, Museum of Decorative Arts

'Gold' is cognate with similar words in many Germanic languages, deriving via Proto-Germanic **gulþą* from Proto-Indo-European **ǵʰelh₃-* ("to shine, to gleam; to be yellow or green").[108][109]

The symbol Au is from the Latin: aurum, the Latin word for "gold".[110] The Proto-Indo-European ancestor of aurum was **h₂é-h₂us-o-*, meaning "glow". This word is derived from the same root (Proto-Indo-European **h₂yes-* "to dawn") as **h₂éysōs*, the ancestor of the Latin word Aurora, "dawn".[111] This etymological relationship is presumably behind the frequent claim in scientific publications that aurum meant "shining dawn".[112]

In popular culture gold is a high standard of excellence, often used in awards.[45] Great achievements are frequently rewarded with gold, in the form of gold medals, gold trophies and other decorations. Winners of athletic events and other graded competitions are usually awarded a gold medal. Many awards such as the Nobel Prize are made from gold as well. Other award statues and prizes are depicted in gold or are gold plated (such as the Academy Awards, the Golden Globe Awards, the Emmy Awards, the Palme d'Or, and the British Academy Film Awards).[113]

Aristotle in his ethics used gold symbolism when referring to what is now known as the golden mean. Similarly, gold is associated with perfect or divine principles, such as in the case of the golden ratio and the golden rule. Gold is further associated with the wisdom of aging and fruition. The fiftieth wedding anniversary is golden. A person's most valued or most successful latter years are sometimes considered "golden years". The height of a civilization is referred to as a golden age.[114]

The first known prehistoric human usages of gold were religious in nature.[115]

In some forms of Christianity and Judaism, gold has been associated both with the sacred and evil. In the Book of Exodus, the Golden Calf is a symbol of idolatry, while in the Book of Genesis, Abraham was said to be rich in gold and silver, and Moses was instructed to cover the Mercy Seat of the Ark of the Covenant with pure gold. In Byzantine iconography the halos of Christ, Virgin Mary and the saints are often golden.[116]

In Islam,[117] gold (along with silk)[118][119] is often cited as being forbidden for men to wear.[120] Abu Bakr al-Jazaeri, quoting a hadith, said that "[t]he wearing of silk and gold are forbidden on the males of my nation, and they are lawful to their women".[121] This, however, has not been enforced consistently throughout history, e.g. in the Ottoman Empire.[122] Further, small gold accents on clothing, such as in embroidery, may be permitted.[123]

In ancient Greek religion and mythology, Theia was seen as the goddess of gold, silver and other gems.[124]

According to Christopher Columbus, those who had something of gold were in possession of something of great value on Earth and a substance to even help souls to paradise.[125]

Wedding rings are typically made of gold. It is long lasting and unaffected by the passage of time and may aid in the ring symbolism of eternal vows before God and the perfection the marriage signifies. In Orthodox Christian wedding ceremonies, the wedded couple is adorned with a golden crown (though some opt for wreaths, instead) during the ceremony, an amalgamation of symbolic rites.[further explanation needed]

On 24 August 2020, Israeli archaeologists discovered a trove of early Islamic gold coins near the central city of Yavne. Analysis of the extremely rare collection of 425 gold coins indicated that they were from the late 9th century. Dating to around 1,100 years back, the gold coins were from the Abbasid Caliphate.[126]

According to the United States Geological Survey in 2016, about 5,726,000,000 troy ounces (178,100 t) of gold has been accounted for, of which 85% remains in active use.[127]

Since the 1880s, South Africa has been the source of a large proportion of the world's gold supply, and about 22% of the gold presently accounted is from South Africa. Production in 1970 accounted for 79% of the world supply, about 1,480 tonnes. In 2007 China (with 276 tonnes) overtook South Africa as the world's largest gold producer, the first time since 1905 that South Africa had not been the largest.[128]

In 2020, China was the world's leading gold-mining country, followed in order by Russia, Australia, the United States, Canada, and Ghana.[9]

In South America, the controversial project Pascua Lama aims at exploitation of rich fields in the high mountains of Atacama Desert, at the border between Chile and Argentina.

It has been estimated that up to one-quarter of the yearly global gold production originates from artisanal or small scale mining.[129][130][131]

The city of Johannesburg located in South Africa was founded as a result of the Witwatersrand Gold Rush which resulted in the discovery of some of the largest natural gold deposits in recorded history. The gold fields are confined to the northern and north-western edges of the Witwatersrand basin, which is a 5-7 km (3.1-4.3 mi) thick layer of Archean rocks located, in most places, deep under the Free State, Gauteng and surrounding provinces.[132] These Witwatersrand rocks are exposed at the surface on the Witwatersrand, in and around Johannesburg, but also in isolated patches to the south-east and south-west of Johannesburg, as well as in an arc around the Vredefort Dome which lies close to the center of the Witwatersrand basin.[65][132] From these surface exposures the basin dips extensively, requiring some of the mining to occur at depths of nearly 4,000 m (13,000 ft), making them, especially the Savuka and TauTona mines to the south-west of Johannesburg, the deepest mines on earth. The gold is found only in six areas where Archean rivers from the north and north-west formed extensive pebbly Braided river deltas before draining into the "Witwatersrand sea" where the rest of the Witwatersrand sediments were deposited.[132]

The Second Boer War of 1899-1901 between the British Empire and the Afrikaner Boers was at least partly over the rights of miners and possession of the gold wealth in South Africa.

During the 19th century, gold rushes occurred whenever large gold deposits were discovered. The first documented discovery of gold in the United States was at the Reed Gold Mine near Georgeville, North Carolina in 1803.[133] The first major gold strike in the United States occurred in a small north Georgia town called Dahlonega.[134] Further gold rushes occurred in California, Colorado, the Black Hills, Otago in New Zealand, a number of locations across Australia, Witwatersrand in South Africa, and the Klondike in Canada.

Grasberg mine located in Papua, Indonesia is the largest gold mine in the world.[135]

Gold extraction is most economical in large, easily mined deposits. Ore grades as little as 0.5 parts per million (ppm) can be economical. Typical ore grades in open-pit mines are 1-5 ppm; ore grades in underground or hard rock mines are usually at least 3 ppm. Because ore grades of 30 ppm are usually needed before gold is visible to the naked eye, in most gold mines the gold is invisible.

The average gold mining and extraction costs were about \$317 per troy ounce in 2007, but these can vary widely depending on mining type and ore quality; global mine production amounted to 2,471.1 tonnes.[136]

After initial production, gold is often subsequently refined industrially by the Wohlwill process which is based on electrolysis or by the Miller process, that is chlorination in the melt. The Wohlwill process results in higher purity, but is more complex and is only applied in small-scale installations.[137][138] Other methods of assaying and purifying smaller amounts of gold include parting and inquartation as well as cupellation, or refining methods based on the dissolution of gold in aqua regia.[139]

In 1997, recycled gold accounted for approximately 20% of the 2700 tons of gold supplied to the market.[140] Jewelry companies such as Generation Collection and computer companies including Dell conduct recycling[141]

As of 2020, the amount of carbon dioxide CO2 produced in mining a kilogram of gold is 16 tonnes, while recycling a kilogram of gold produces 53 kilograms of CO2 equivalent. Approximately 30 percent of the global gold supply is recycled and not mined as of 2020.[142]

The consumption of gold produced in the world is about 50% in jewelry, 40% in investments, and 10% in industry.[11][146]

According to the World Gold Council, China was the world's largest single consumer of gold in 2013, overtaking India.[147]

Gold production is associated with contribution to hazardous pollution.[148][149]

Low-grade gold ore may contain less than one ppm gold metal; such ore is ground and mixed with sodium cyanide to dissolve the gold. Cyanide is a highly poisonous chemical, which can kill living creatures when exposed in minute quantities. Many cyanide spills[150] from gold mines have occurred in both developed and developing countries which killed aquatic life in long stretches of affected rivers. Environmentalists consider these events major environmental disasters.[151][152] Up to thirty tons of used ore can be dumped as waste for producing one troy ounce of gold.[153] Gold ore dumps are the source of many heavy elements such as cadmium, lead, zinc, copper, arsenic, selenium and mercury. When sulfide-bearing minerals in these ore dumps are exposed to air and water, the sulfide transforms into sulfuric acid which in turn dissolves these heavy metals facilitating their passage into surface water and ground water. This process is called acid mine drainage. These gold ore dumps contain long-term, highly hazardous waste.[153]

It was once common to use mercury to recover gold from ore, but today the use of mercury is largely limited to small-scale individual miners.[154] Minute quantities of mercury compounds can reach water bodies, causing heavy metal contamination. Mercury can then enter into the human food chain in the form of methylmercury. Mercury poisoning in humans causes incurable brain function damage and severe retardation.[155]

Gold extraction is also a highly energy-intensive industry, extracting ore from deep mines and grinding the large quantity of ore for further chemical extraction requires nearly 25 kWh of electricity per gram of gold produced.[156]

Gold has been widely used throughout the world as money,[157] for efficient indirect exchange (versus barter), and to store wealth in hoards. For exchange purposes, mints produce standardized gold bullion coins, bars and other units of fixed weight and purity.

The first known coins containing gold were struck in Lydia, Asia Minor, around 600 BC.[90] The talent coin of gold in use during the periods of Grecian history both before and during the time of the life of Homer weighed between 8.42 and 8.75 grams.[158] From an earlier preference in using silver, European economies re-established the minting of gold as coinage during the thirteenth and fourteenth centuries.[159]

Bills (that mature into gold coin) and gold certificates (convertible into gold coin at the issuing bank) added to the circulating stock of gold standard money in most 19th century industrial economies. In preparation for World War I the warring nations moved to fractional gold standards, inflating their currencies to finance the war effort. Post-war, the victorious countries, most notably Britain, gradually restored gold-convertibility, but international flows of gold via bills of exchange remained embargoed; international shipments were made exclusively for bilateral trades or to pay war reparations.

After World War II gold was replaced by a system of nominally convertible currencies related by fixed exchange rates following the Bretton Woods system. Gold standards and the direct convertibility of currencies to gold have been abandoned by world governments, led in 1971 by the United States' refusal to redeem its dollars in gold. Fiat currency now fills most monetary roles. Switzerland was the last country to tie its currency to gold; this was ended by a referendum in 1999.[160]

Central banks continue to keep a portion of their liquid reserves as gold in some form, and metals exchanges such as the London Bullion Market Association still clear transactions denominated in gold, including future delivery contracts. Today, gold mining output is declining.[161] With the sharp growth of economies in the 20th century, and increasing foreign exchange, the world's gold reserves and their trading market have become a small fraction of all markets and fixed exchange rates of currencies to gold have been replaced by floating prices for gold and gold future contract. Though the gold stock grows by only 1% or 2% per year, very little metal is irretrievably consumed. Inventory above ground would satisfy many decades of industrial and even artisan uses at current prices.

The gold proportion (fineness) of alloys is measured by karat (k). Pure gold (commercially termed fine gold) is designated as 24 karat, abbreviated 24k. English gold coins intended for circulation from 1526 into the 1930s were typically a standard 22k alloy called crown gold,[162] for hardness (American gold coins for circulation after 1837 contain an alloy of 0.900 fine gold, or 21.6 kt).[163]

Although the prices of some platinum group metals can be much higher, gold has long been considered the most desirable of precious metals, and its value has been used as the standard for many currencies. Gold has been used as a symbol for purity, value, royalty, and particularly roles that combine these properties. Gold as a sign of wealth and prestige was ridiculed by Thomas More in his treatise Utopia. On that imaginary island, gold is so abundant that it is used to make chains for slaves, tableware, and lavatory seats. When ambassadors from other countries arrive, dressed in ostentatious gold jewels and badges, the Utopians mistake them for menial servants, paying homage instead to the most modestly dressed of their party.

The ISO 4217 currency code of gold is XAU.[164] Many holders of gold store it in form of bullion coins or bars as a hedge against inflation or other economic disruptions, though its efficacy as such has been questioned; historically, it has not proven itself reliable as a hedging instrument.[165] Modern bullion coins for investment or collector purposes do not require good mechanical wear properties; they are typically fine gold at 24k, although the American Gold Eagle and the British gold sovereign continue to be minted in 22k (0.92) metal in historical tradition, and the South African Krugerrand, first released in 1967, is also 22k (0.92).[166]

The special issue Canadian Gold Maple Leaf coin contains the highest purity gold of any bullion coin, at 99.999% or 0.99999, while the popular issue Canadian Gold Maple Leaf coin has a purity of 99.99%. In 2006, the United States Mint began producing the American Buffalo gold bullion coin with a purity of 99.99%. The Australian Gold Kangaroos were first coined in 1986 as the Australian Gold Nugget but changed the reverse design in 1989. Other modern coins include the Austrian Vienna Philharmonic bullion coin and the Chinese Gold Panda.[167]

Like other precious metals, gold is measured by troy weight and by grams. The proportion of gold in the alloy is measured by karat (k), with 24 karat (24k) being pure gold (100%), and lower karat numbers proportionally less (18k = 75%). The purity of a gold bar or coin can also be expressed as a decimal figure ranging from 0 to 1, known as the millesimal fineness, such as 0.995 being nearly pure.

The price of gold is determined through trading in the gold and derivatives markets, but a procedure known as the Gold Fixing in London, originating in September 1919, provides a daily benchmark price to the industry. The afternoon fixing was introduced in 1968 to provide a price when US markets are open.[168] As of September 2017[update], gold was valued at around \$42 per gram (\$1,300 per troy ounce).

Historically gold coinage was widely used as currency; when paper money was introduced, it typically was a receipt redeemable for gold coin or bullion. In a monetary system known as the gold standard, a certain weight of gold was given the name of a unit of currency. For a long period, the United States government set the value of the US dollar so that one troy ounce was equal to \$20.67 (\$0.665 per gram), but in 1934 the dollar was devalued to \$35.00 per troy ounce (\$0.889/g). By 1961, it was becoming hard to maintain this price, and a pool of US and European banks agreed to manipulate the market to prevent further currency devaluation against increased gold demand.[169]

The largest gold depository in the world is that of the U.S. Federal Reserve Bank in New York, which holds about 3%[170] of the gold known to exist and accounted for today, as does the similarly laden U.S. Bullion Depository at Fort Knox. In 2005 the World Gold Council estimated total global gold supply to be 3,859 tonnes and demand to be 3,754 tonnes, giving a surplus of 105 tonnes.[171]

After 15 August 1971 Nixon shock, the price began to greatly increase,[172] and between 1968 and 2000 the price of gold ranged widely, from a high of \$850 per troy ounce (\$27.33/g) on 21 January 1980, to a low of \$252.90 per troy ounce (\$8.13/g) on 21 June 1999 (London Gold Fixing).[173] Prices increased rapidly from 2001, but the 1980 high was not exceeded until 3 January 2008, when a new maximum of \$865.35 per troy ounce was set.[174] Another record price was set on 17 March 2008, at \$1023.50 per troy ounce (\$32.91/g).[174]

On 2 December 2009, gold reached a new high closing at \$1,217.23.[175] Gold further rallied hitting new highs in May 2010 after the European Union debt crisis prompted further purchase of gold as a safe asset.[176][177] On 1 March 2011, gold hit a new all-time high of \$1432.57, based on investor concerns regarding ongoing unrest in North Africa as well as in the Middle East.[178]

From April 2001 to August 2011, spot gold prices more than quintupled in value against the US dollar, hitting a new all-time high of \$1,913.50 on 23 August 2011,[179] prompting speculation that the long secular bear market had ended and a bull market had returned.[180] However, the price then began a slow decline towards \$1200 per troy ounce in late 2014 and 2015.

In August 2020, the gold price picked up to US\$2060 per ounce after a total growth of 59% from August 2018 to October 2020, a period during which it outpaced the Nasdaq total return of 54%.[181]

Gold futures are traded on the COMEX exchange.[182] These contracts are priced in USD per troy ounce (1 troy ounce = 31.1034768 grams).[183] Below are the CQG contract specifications outlining the futures contracts:

Because of the softness of pure (24k) gold, it is usually alloyed with other metals for use in jewelry, altering its hardness and ductility, melting point, color and other properties. Alloys with lower karat rating, typically 22k, 18k, 14k or 10k, contain higher percentages of copper, silver, palladium or other base metals in the alloy.[24] Nickel is toxic, and its release from nickel white gold is controlled by legislation in Europe.[24] Palladium-gold alloys are more expensive than those using nickel.[184] High-karat white gold alloys are more resistant to corrosion than are either pure silver or sterling silver. The Japanese craft of Mokume-gane exploits the color contrasts between laminated colored gold alloys to produce decorative wood-grain effects.

By 2014, the gold jewelry industry was escalating despite a dip in gold prices. Demand in the first quarter of 2014 pushed turnover to \$23.7 billion according to a World Gold Council report.

Gold solder is used for joining the components of gold jewelry by high-temperature hard soldering or brazing. If the work is to be of hallmarking quality, the gold solder alloy must match the fineness of the work, and alloy formulas are manufactured to color-match yellow and white gold. Gold solder is usually made in at least three melting-point ranges referred to as Easy, Medium and Hard. By using the hard, high-melting point solder first, followed by solders with progressively lower melting points, goldsmiths can assemble complex items with several separate soldered joints. Gold can also be made into thread and used in embroidery.

Only 10% of the world consumption of new gold produced goes to industry,[11] but by far the most important industrial use for new gold is in fabrication of corrosion-free electrical connectors in computers and other electrical devices. For example, according to the World Gold Council, a typical cell phone may contain 50 mg of gold, worth about 2 dollars 82 cents. But since nearly one billion cell phones are produced each year, a gold value of US\$2.82 in each phone adds to US\$2.82 billion in gold from just this application.[185] (Prices updated to November 2022)

Though gold is attacked by free chlorine, its good conductivity and general resistance to oxidation and corrosion in other environments (including resistance to non-chlorinated acids) has led to its widespread industrial use in the electronic era as a thin-layer coating on electrical connectors, thereby ensuring good connection. For example, gold is used in the connectors of the more expensive electronics cables, such as audio, video and USB cables. The benefit of using gold over other connector metals such as tin in these applications has been debated; gold connectors are often criticized by audio-visual experts as unnecessary for most consumers and seen as simply a marketing ploy. However, the use of gold in other applications in electronic sliding contacts in highly humid or corrosive atmospheres, and in use for contacts with a very high failure cost (certain computers, communications equipment, spacecraft, jet aircraft engines) remains very common.[186]

Besides sliding electrical contacts, gold is also used in electrical contacts because of its resistance to corrosion, electrical conductivity, ductility and lack of toxicity.[187] Switch contacts are generally subjected to more intense corrosion stress than are sliding contacts. Fine gold wires are used to connect semiconductor devices to their packages through a process

s known as wire bonding.

The concentration of free electrons in gold metal is $5.91 \times 10^{22} \text{ cm}^{-3}$.^[188] Gold is highly conductive to electricity and has been used for electrical wiring in some high-energy applications (only silver and copper are more conductive per volume, but gold has the advantage of corrosion resistance). For example, gold electrical wires were used during some of the Manhattan Project's atomic experiments, but large high-current silver wires were used in the calutron isotope separator magnets in the project.

It is estimated that 16% of the world's presently-accounted-for gold and 22% of the world's silver is contained in electronic technology in Japan.^[189]

Metallic and gold compounds have long been used for medicinal purposes. Gold, usually as the metal, is perhaps the most anciently administered medicine (apparently by shamanic practitioners)^[190] and known to Dioscorides.^{[191][192]} In medieval times, gold was often seen as beneficial for the health, in the belief that something so rare and beautiful could not be anything but healthy. Even some modern esotericists and forms of alternative medicine assign metallic gold a healing power.

In the 19th century gold had a reputation as an anxiolytic, a therapy for nervous disorders. Depression, epilepsy, migraine, and glandular problems such as amenorrhea and impotence were treated, and most notably alcoholism (Keeley, 1897).^[193]

The apparent paradox of the actual toxicology of the substance suggests the possibility of serious gaps in the understanding of the action of gold in physiology.^[194] Only salts and radioisotopes of gold are of pharmacological value, since elemental (metallic) gold is inert to all chemicals it encounters inside the body (e.g., ingested gold cannot be attacked by stomach acid). Some gold salts do have anti-inflammatory properties and at present two are still used as pharmaceuticals in the treatment of arthritis and other similar conditions in the US (sodium aurothiomalate and auranofin). These drugs have been explored as a means to help to reduce the pain and swelling of rheumatoid arthritis, and also (historically) against tuberculosis and some parasites.^[195]

Gold alloys are used in restorative dentistry, especially in tooth restorations, such as crowns and permanent bridges. The gold alloys' slight malleability facilitates the creation of a superior molar mating surface with other teeth and produces results that are generally more satisfactory than those produced by the creation of porcelain crowns. The use of gold crowns is more prominent teeth such as incisors is favored in some cultures and discouraged in others.

Colloidal gold preparations (suspensions of gold nanoparticles) in water are intensely red-colored, and can be made with tightly controlled particle sizes up to a few tens of nanometers across by reduction of gold chloride with citrate or ascorbate ions. Colloidal gold is used in research applications in medicine, biology and materials science. The technique of immunogold labeling exploits the ability of the gold particles to adsorb protein molecules onto their surfaces. Colloidal gold particles coated with specific antibodies can be used as probes for the presence and position of antigens on the surfaces of cells.^[196] In ultrathin sections of tissues viewed by electron microscopy, the immunogold labels appear as extremely dense round spots at the position of the antigen.^[197]

Gold, or alloys of gold and palladium, are applied as conductive coating to biological specimens and other non-conducting materials such as plastics and glass to be viewed in a scanning electron microscope. The coating, which is usually applied by sputtering with an argon plasma, has a triple role in this application. Gold's very high electrical conductivity drains electrical charge to earth, and its very high density provides stopping power for electrons in the electron beam, helping to limit the depth to which the electron beam penetrates the specimen. This improves definition of the position and topography of the specimen surface and increases the spatial resolution of the image. Gold also produces a high output of secondary electrons when irradiated by an electron beam, and these low-energy electrons are the most commonly used signal source used in the scanning electron microscope.^[198]

The isotope gold-198 (half-life 2.7 days) is used in nuclear medicine, in some cancer treatments and for treating other diseases.^{[199][200]}

Pure metallic (elemental) gold is non-toxic and non-irritating when ingested^[213] and is sometimes used as a food decoration in the form of gold leaf.^[214] Metallic gold is also a component of the alcoholic drinks Goldschläger, Gold Strike, and Goldwasser. Metallic gold is approved as a food additive in the EU (E175 in the Codex Alimentarius). Although the gold ion is toxic, the acceptance of metallic gold as a food additive is due to its relative chemical inertness, and resistance to being corroded or transformed into soluble salts (gold compounds) by any known chemical process which would be encountered in the human body.

Soluble compounds (gold salts) such as gold chloride are toxic to the liver and kidneys. Common cyanide salts of gold such as potassium gold cyanide, used in gold electroplating, are toxic by virtue of both their cyanide and gold content. There are rare cases of lethal gold poisoning from potassium gold cyanide.^{[215][216]} Gold toxicity can be ameliorated with chelation therapy with an agent such as dimercaprol.

Gold metal was voted Allergen of the Year in 2001 by the American Contact Dermatitis Society; gold contact allergies affect mostly women.^[217] Despite this, gold is a relatively non-potent contact allergen, in comparison with metals like nickel.^[218]

A sample of the fungus *Aspergillus niger* was found growing from gold mining solution; and was found to contain cyanometal complexes, such as gold, silver, copper, iron and zinc. The fungus also plays a role in the solubilization of heavy metal sulfides.^[219]

RegEx

Now we have a large string with all the text, but you can see that there are blank new lines, so we will filter those out using the RegEx library, which is similar to the re library but with more robust features.

```
In [30]: text_no_newline = regex.sub('\n', '', all_text)

print(text_no_newline)
```

ers.[151][152] Up to thirty tons of used ore can be dumped as waste for producing one troy ounce of gold.[153] Gold ore dumps are the source of many heavy elements such as cadmium, lead, zinc, copper, arsenic, selenium and mercury. When sulfide-bearing minerals in these ore dumps are exposed to air and water, the sulfide transforms into sulfuric acid which in turn dissolves these heavy metals facilitating their passage into surface water and ground water. This process is called acid mine drainage. These gold ore dumps contain long-term, highly hazardous waste.[153] It was once common to use mercury to recover gold from ore, but today the use of mercury is largely limited to small-scale individual miners.[154] Minute quantities of mercury compounds can reach water bodies, causing heavy metal contamination. Mercury can then enter into the human food chain in the form of methylmercury. Mercury poisoning in humans causes incurable brain function damage and severe retardation.[155] Gold extraction is also a highly energy-intensive industry, extracting ore from deep mines and grinding the large quantity of ore for further chemical extraction requires nearly 25 kWh of electricity per gram of gold produced.[156] Gold has been widely used throughout the world as money,[157] for efficient indirect exchange (versus barter), and to store wealth in hoards. For exchange purposes, mints produce standardized gold bullion coins, bars and other units of fixed weight and purity. The first known coins containing gold were struck in Lydia, Asia Minor, around 600 BC.[90] The talent coin of gold in use during the periods of Grecian history both before and during the time of the life of Homer weighed between 8.42 and 8.75 grams.[158] From an earlier preference in using silver, European economies re-established the minting of gold as coinage during the thirteenth and fourteenth centuries.[159] Bills (that mature into gold coin) and gold certificates (convertible into gold coin at the issuing bank) added to the circulating stock of gold standard money in most 19th century industrial economies. In preparation for World War I the warring nations moved to fractional gold standards, inflating their currencies to finance the war effort. Post-war, the victorious countries, most notably Britain, gradually restored gold-convertibility, but international flows of gold via bills of exchange remained embargoed; international

Tokenization

Now that the new lines are gone, we are going to do a basic `.strip().split('.')` to tokenize the long string into sentences using the periods as splitting points.

```
In [31]: # tokenize using strip() and split()
sentences = text_no_newline.strip().split('.')

print(sentences)
```

['Gold is a chemical element; it has symbol Au (from Latin aurum\xa0gold) and atomic number 79', ' In pure form, it is a bright, slightly orange-yellow, dense, soft, malleable, and ductile metal', ' Chemically, gold is a transition metal, a group 11 element, and one of the noble metals', ' It is one of the least reactive chemical elements and is solid under standard conditions', 'Gold often occurs in free elemental (native state), as nuggets or grains, in rocks, veins, and alluvial deposits', ' It occurs in a solid solution series with the native element silver (as in electrum), naturally alloyed with other metals like copper and palladium, and mineral inclusions such as within pyrite', ' Less commonly, it occurs in minerals as gold compounds, often with tellurium (gold tellurides)', 'Gold is resistant to most acids, though it does dissolve in aqua regia (a mixture of nitric acid and hydrochloric acid), forming a soluble tetrachloroaurate anion', ' Gold is insoluble in nitric acid alone, which dissolves silver and base metals, a property long used to refine gold and confirm the presence of gold in metallic substances, giving rise to the term acid test', ' Gold dissolves in alkaline solutions of cyanide, which are used in mining and electroplating', ' Gold also dissolves in mercury, forming amalgam alloys, and as the gold acts simply as a solute, this is not a chemical reaction', 'A relatively rare element,[7][8] gold is a precious metal that has been used for coinage, jewelry, and other arts throughout recorded history', ' In the past, a gold standard was often implemented as a monetary policy', ' Gold coins ceased to be minted as a circulating currency in the 1930s, and the world gold standard was abandoned for a fiat currency system after the Nixon shock measures of 1971', 'In 2020, the world's largest gold producer was China, followed by Russia and Australia', '[9] A total of around 201,296 tonnes of gold exists above ground, as of 2020[update]', '[10] This is equal to a cube with each side measuring roughly 21', '7 meters (71\xa0ft)', ' The world's consumption of new gold produced is about 50% in jewelry, 40% in investments, and 10% in industry', '[11] Gold's high malleability, ductility, resistance to corrosion and most other chemical reactions, and conductivity

```
In [26]: len(sentences)
```

Out[26]: 465

Please note that this removed the period for whatever reason, so we must put it back when we join.

Alternatively, we can also use `sent_tokenize` to tokenize the text into sentences. However, this gives us less sentences after tokenization. We can visually see that sentences that end in a period but with no space after the period were not properly tokenized.

```
In [34]: # tokenize into sentences using sent_tokenize
sent_tokenized = sent_tokenize(text_no_newline)
for i in range(len(sent_tokenized)):
    sent_tokenized[i] = regex.sub('\'', '', sent_tokenized[i])
    sent_tokenized[i] = regex.sub('\\"', '', sent_tokenized[i])
print(len(sent_tokenized))
print(sent_tokenized)
```

355

['Gold is a chemical element; it has symbol Au (from Latin aurum\xa0gold) and atomic number 79.', 'In pure form, it is a bright, slightly orange-yellow, dense, soft, malleable, and ductile metal.', 'Chemically, gold is a transition metal, a group 11 element, and one of the noble metals.', 'It is one of the least reactive chemical elements and is solid under standard conditions. Gold often occurs in free elemental (native state), as nuggets or grains, in rocks, veins, and alluvial deposits.', 'It occurs in a solid solution series with the native element silver (as in electrum), naturally alloyed with other metals like copper and palladium, and mineral inclusions such as within pyrite.', 'Less commonly, it occurs in minerals as gold compounds, often with tellurium (gold tellurides). Gold is resistant to most acids, though it does dissolve in aqua regia (a mixture of nitric acid and hydrochloric acid), forming a soluble tetrachloroaurate anion.', 'Gold is insoluble in nitric acid alone, which dissolves silver and base metals, a property long used to refine gold and confirm the presence of gold in metallic substances, giving rise to the term acid test.', 'Gold dissolves in alkaline solutions of cyanide, which are used in mining and electroplating.', 'Gold also dissolves in mercury, forming amalgam alloys, and as the gold acts simply as a solute, this is not a chemical reaction. A relatively rare element, [7][8] gold is a precious metal that has been used for coinage, jewelry, and other arts throughout recorded history.', 'In the past, a gold standard was often implemented as a monetary policy.', 'Gold coins ceased to be minted as a circulating currency in the 1930s, and the world gold standard was abandoned for a fiat currency system after the Nixon shock measures of 1971. In 2020, the world's largest gold producer was China, followed by Russia and Australia.', '[9] A total of around 201,296 tonnes of gold exists above ground, as of 2020[update].', '[10] This is equal to a cube with each side measuring roughly 21.7 meters (71\xa0ft).', 'The world's consumption of new gold produced is about 50% in jewelry, 40% in investments, and 10% in industry.',

To move forward we will go with using the strip().split() method since it seems some of the sentences did not get split using the sent_tokenize() method.

EDA (Data cleaning and prep)

Next, we will clean the text some more so we can use in our pretrained models. Of course, we will start by removing stopwords, punctuations, and create a list of all the words after cleaning.

```
In [36]: nltk.download("punkt")
nltk.download("stopwords")
```

```
[nltk_data] Downloading package punkt to
[nltk_data]   C:\Users\etran\AppData\Roaming\nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]   C:\Users\etran\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
```

Out[36]: True

```
In [75]: # This function takes in a string and returns tokenized sentences and words
def preprocess_text(text):
    """sentences = sent_tokenize(text) # tokenize into sentences
    stop_words = set(stopwords.words("english")) # removes stop words

    words = [word_tokenize(sentence.lower()) for sentence in sentences] # tokenize into words that are lower case
    words = [[word for word in sentence if word.isalnum() and word not in stop_words] for sentence in words] # removes special characters
    """
    stop_words = set(stopwords.words("english")) # removes stop words
    text_no_newline = regex.sub('\n', '', all_text)
    sentences = text_no_newline.strip().split('.')
    #words = [word_tokenize(sentence.lower()) for sentence in sentences] # tokenize into words that are lower case
    #words = [[word for word in sentence if word.isalnum() and word not in stop_words] for sentence in words] # removes special characters

    cleaned_words = []
    for idx_sent in range(len(sentences)):
        temp = sentences[idx_sent].lower().split()
        for idx_word in range(len(temp)):
            if temp[idx_word].isalnum() and temp[idx_word] not in stop_words:
                cleaned_words.append(temp[idx_word])
    #cleaned_words.append = [word.lower() for word in cleaned_words]
    return sentences, cleaned_words
```

We will now run the function we just made, passing in our long string. We will get out cleaned sentences and cleaned words.


```
In [76]: processed_sentences, processed_words = preprocess_text(all_text)
print(len(processed_sentences))
print(processed_sentences)
```

meis where he gave away so much gold that it depressed the price in Egypt for over a decade, causing high inflation', '[91] A contemporary Arab historian remarked:Gold was at a high price in Egypt until they came in that year', 'The mithqal did not go below 25 dirhams and was generally above, but from that time its value fell and it cheapened in price and has remained cheap till now', 'The mithqal does not exceed 22 dirhams or less', 'This has been the state of affairs for about twelve years until this day by reason of the large amount of gold which they brought into Egypt and spent there ['', '', ']', 'The European exploration of the Americas was fueled in no small part by reports of the gold ornaments displayed in great profusion by Native American peoples, especially in Mesoamerica, Peru, Ecuador and Colombia', 'The Aztecs regarded gold as the product of the gods, calling it literally god excrement (teocuitlatl in Nahuatl), and after Moctezuma II was killed, most of this gold was shipped to Spain', '[94] However, for the indigenous peoples of North America gold was considered useless and they saw much greater value in other minerals which were directly related to their utility, such as obsidian, flint, and slate', '[95] El Dorado is applied to a legendary story in which precious stones were found in fabulous abundance along with gold coins', 'The concept of El Dorado underwent several transformations, and eventually accounts of the previous myth were also combined with those of a legendary lost city', 'El Dorado, was the term used by the Spanish Empire to describe a mythical tribal chief (zipa) of the Muisca native people in Colombia, who, as an initiation rite, covered himself with gold dust and submerged in Lake Guatavita', 'The legends surrounding El Dorado changed over time, as it went from being a man, to a city, to a kingdom, and then finally to an empire', 'Beginning in the early modern period, European exploration and colonization of West Africa was driven in large part by reports of gold deposits in the region, which was eventually referred to by Europeans as the Gold Coast', '[96] From the late 15th to early 19th centuries, European trade in the region was primarily focused in gold, along with ivory and slaves', '[97] The gold trade in West Africa was dominated by the Ashanti Empire, who initially traded with the Portuguese before branching out and trading

```
In [77]: print(len(processed_words))
print(processed_words)
```

5890
['gold', 'chemical', 'element;', 'symbol', 'au', '(from', 'latin', 'aurum', 'gold)', 'atomic', 'number', '79', 'pure', 'form', 'bright', 'slightly', 'orange-yellow', 'dense', 'soft', 'malleable', 'ductile', 'metal', 'chemically', 'gold', 'transition', 'metal', 'group', '11', 'element', 'one', 'noble', 'metals', 'one', 'least', 'reactive', 'chemical', 'elements', 'solid', 'standard', 'conditions', 'gold', 'often', 'occurs', 'free', 'elemental', '(native', 'state)', 'nuggets', 'grains', 'rocks', 'veins', 'alluvial', 'deposits', 'occurs', 'solid', 'solution', 'series', 'native', 'element', 'silver', '(as', 'electrum)', 'naturally', 'alloyed', 'metals', 'like', 'copper', 'palladium', 'mineral', 'inclusions', 'within', 'pyrite', 'less', 'commonly', 'occurs', 'minerals', 'gold', 'compounds', 'often', 'tellurium', '(gold', 'tellurides', 'gold', 'resistant', 'acids', 'though', 'dissolve', 'aqua', 'regia', '(a', 'mixture', 'nitric', 'acid', 'hydrochloric', 'acid)', 'forming', 'soluble', 'tetrachloroaurate', 'anion', 'gold', 'insoluble', 'nitric', 'acid', 'alone', 'dissolves', 'silver', 'base', 'metals', 'property', 'long', 'used', 'refine', 'gold', 'confirm', 'presence', 'gold', 'metallic', 'substances', 'giving', 'rise', 'term', 'acid', 'test', 'gold', 'dissolves', 'alkaline', 'solutions', 'cyanide', 'used', 'mining', 'electroplating', 'gold', 'also', 'dissolves', 'mercury', 'forming', 'amalgam', 'alloys', 'gold', 'acts', 'simply', 'solute', 'chemical', 'reaction', 'relatively', 'rare', 'element', '[7][8]', 'gold', 'precious', 'metal', 'used', 'coinage', 'jewelry', 'arts', 'throughout', 'recorded', 'history', 'past', 'gold', 'standard', 'often', 'implemented', 'monetary', 'policy', 'gold', 'coins', 'ceased', 'minted', 'circulating', 'currency', '1930s', 'world', 'gold', 'standard', 'abandoned', 'fiat', 'currency', 'system', 'nixon', 'shock', 'measures', '1971', '2020', 'world', 'largest', 'gold', 'producer', 'china', 'followed', 'russia', 'australia', '[9]', 'total', 'around', '201,296', 'tonnes', 'gold', 'exists', 'ground', '2020[update]', '[10]', 'equal', 'cube', 'side', 'measuring', 'roughly', '21', '7', 'm

Now it's time to start getting useful information out of our data using some of the libraries. We will start with nltk to find the parts of speech (POS), then we will use transformers, specifically the pipeline module that contains many pretrained models.

POS

Parts of speech is very important in NLP. We will use the nltk library to tag each word.

```
In [107]: nltk.pos_tag(processed_words) # tags each word with its part of speech tag
```

```
(('orange-yellow', 'JJ'),
 ('dense', 'NN'),
 ('soft', 'NN'),
 ('malleable', 'NN'),
 ('ductile', 'IN'),
 ('metal', 'NN'),
 ('chemically', 'NN'),
 ('gold', 'NN'),
 ('transition', 'NN'),
 ('metal', 'NN'),
 ('group', 'NN'),
 ('11', 'CD'),
 ('element', 'NN'),
 ('one', 'CD'),
 ('noble', 'JJ'),
 ('metals', 'NNS'),
 ('one', 'CD'),
 ('least', 'JJ'),
 ('reactive', 'JJ'),
 ('chemical', 'JJ'),
```

Transformers

```
In [102]: ► import transformers # library with pipeline
transformers.logging.set_verbosity_error() # to display errors

from transformers import pipeline # class with many pretrained models for various NLP tasks, from transformers library
```

The first thing we are going to do with the pipeline is make an instance using the NER task setting, which will tag each entity with its NER tag.

```
In [104]: ► nlp = pipeline(task='ner') # this creates instance of pipeline class that is trained with NER functionality
for item in nlp(all_text): # iterates thru each word and performs Named Entity Recognition and tagging
    print(f"{item['word']}, {item['entity']}") # prints NER results of each entity found

('Latin', 'I-MISC')
('Nixon', 'I-PER')
('China', 'I-LOC')
('Russia', 'I-LOC')
('Australia', 'I-LOC')
```

Of course, we can also do NER using SpaCy and displacy to visualize it, which looks much better.

```
In [106]: ► spacy_nlp = spacy.load('en_core_web_sm')
doc = spacy_nlp(all_text)
displacy.render(doc, style='ent', jupyter=True)
```

Gold is a chemical element; it has symbol Au **PERSON** (from Latin **LANGUAGE** aurum 'gold') and atomic number 79 **CARDINAL**. In pure form, it is a bright, slightly orange-yellow, dense, soft, malleable, and ductile metal. Chemically, gold is a transition metal, a group 11 **CARDINAL** element, and one **CARDINAL** of the noble metals. It is one of the least reactive chemical elements and is solid under standard conditions. Gold often occurs in free elemental (native state), as nuggets or grains, in rocks, veins, and alluvial deposits. It occurs in a solid solution series with the native element silver (as in electrum), naturally alloyed with other metals like copper and palladium, and mineral inclusions such as within pyrite. Less commonly, it occurs in minerals as gold compounds, often with tellurium (gold tellurides). Gold is resistant to most acids, though it does dissolve in aqua regia (a mixture of nitric acid and hydrochloric acid), forming a soluble tetrachloroaurate anion. Gold is insoluble in nitric acid alone, which dissolves silver and base metals, a property long used to refine gold and confirm the presence of

In []: ►

In []: ►

In []: ►

Next, we will make another instance of the pipeline, but this time the task will be set to summarization. With this, we can easily get the summary of the article.

```
In [108]: # create a summarization pipeline using a pre-trained model designed for text summarization.
summary_pipe = pipeline(task='summarization')
#sum_len = len(all_text)/20 # this variable is 5% of the number of characters in the document
sum_len = 500
# uses the nlp function to perform text summarization on the input text stored in the txt variable.
summary = summary_pipe(all_text, max_length=sum_len, min_length=30)
print(summary)
```

ResourceExhaustedError: Exception encountered when calling layer 'SelfAttention' (type TFT5Attention).

```
{{function_node __wrapped__Einsum_N_2_device_/job:localhost/replica:0/task:0/device:CPU:0}} OOM when allocating tensor with shape[1,8,15278,15278] and type float on /job:localhost/replica:0/task:0/device:CPU:0 by allocator cpu [Op:Einsum] name:
e:
```

Call arguments received by layer 'SelfAttention' (type TFT5Attention):

- `hidden_states=tf.Tensor(shape=(1, 15278, 512), dtype=float32)`
- `mask=tf.Tensor(shape=(1, 1, 1, 15278), dtype=float32)`
- `key_value_states=None`
- `position_bias=None`
- `past_key_value=None`
- `layer_head_mask=None`
- `query_length=None`
- `use_cache=False`
- `training=False`
- `output_attentions=False`

Sentiment Analysis

We can perform sentiment analysis on the summary by again using the pipeline pretrained models. For this we will use 'text_classification' as the task type.

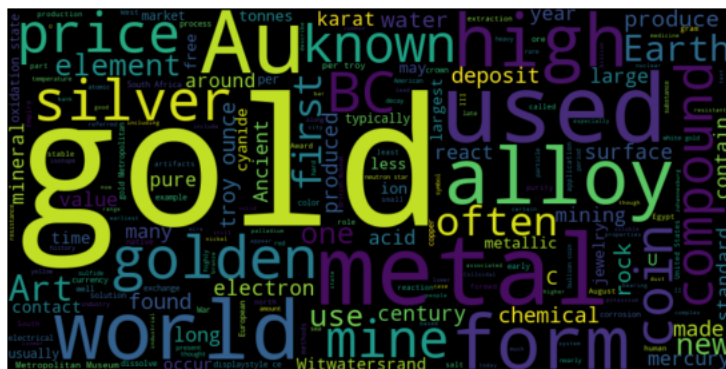
```
In [ ]: sa_pipe = pipeline("text-classification") # instantiate a pretrained BERT model for sentiment analysis
```

```
In [ ]: sa_pipe(summary)
```

Reading the summary, it seems like an efficient way to summarize the content as these sentences seem very relevant and informative.

Next, let's do some visualization. Here is a basic wordcloud using the original text.

```
In [109]: wordcloud = WordCloud(width=800, height=400, background_color="black").generate(all_text)
plt.figure(figsize=(10,5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



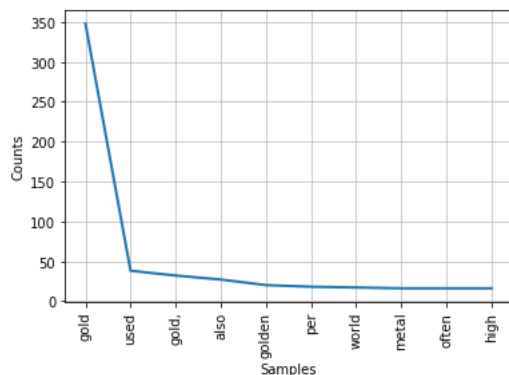
We will continue with the visualization by using `FreqDist`, a module in `nltk` that lets us find the most common words, as well as plot them easily.

```
In [110]: from nltk.probability import FreqDist
          fdist = FreqDist(processed_words)
```

```
In [111]: fdist.most_common(10)
```

```
Out[111]: [('gold', 348),  
           ('used', 38),  
           ('gold', 32),  
           ('also', 27),  
           ('golden', 20),  
           ('per', 18),  
           ('world', 17),  
           ('metal', 16),  
           ('often', 16),  
           ('high', 16)]
```

```
In [112]: fdist.plot(10) # plots the 10 most common words
```



```
Out[112]: <AxesSubplot:xlabel='Samples', ylabel='Counts'>
```

Conclusions

As we can see from this project, there are many tools available to us today to easily perform some of the NLP tasks that once took lots and lots of code. The new tools are more powerful, although they don't always work better as was the case with the `sent_tokenize` and `word_tokenize` modules. They also seemed to take more time and processing, as was seen from using pipeline vs spacy for NER.

Also, my summary did not work on my machine due to memory issues, and the pipeline module could not load on the virtual machines. This is probably due to path and transformers versions. I ran out of time and did not get to the chatbot portion of the project, although I have everything else complete other than the chatbot. I will revisit this at a later date to fix the pipeline/transformers, enter the chatbot, add the necessary modules, and update to my GitHub repository.

```
In [113]: # DO NOT RUN DOES NOT WORK

"""import gensim
from gensim.models import Word2Vec
# from gensim.summarization import summarize

model = Word2Vec(processed_words, min_count=1, vector_size=100, window=5, sg=1)
model.build_vocab(processed_words)
model.train(processed_words, total_examples=model.corpus_count, epochs=20)
sentence_vectors = []
for sentence in processed_words:
    if len(sentence) != 0:
        sentence_vec = sum([model.wv[word] for word in sentence]) #/ len(sentence)
        sentence_vectors.append(sentence_vec)

#similarity_matrix = [[model.wv.similarity(w1, w2) for w1 in sentence_vectors] for w2 in sentence_vectors]
#similarity_scores = similarity_matrix.sum(axis=1) # axis = 1 means horizontal
ms = model.wv.most_similar('gold', topn= 6)[1:]
#top_sentences_indices = np.argsort(ms, axis=0)[-5:] # axis = 0 means sort vertically, and get the top 5
#top_sentences = []
for index in range(len(ms)):
    val = processed_sentences[index]
    summary = ''.join(val) + '.'
    #print(summary)

#summary = summarize(" ".join(sentences), ratio=0.3, word_count=None, split=False)

"""
```

```
Out[113]: 'import gensim\nfrom gensim.models import Word2Vec\n# from gensim.summarization import summarize\n\nmodel = Word2Vec(proce\nssed_words, min_count=1, vector_size=100, window=5, sg=1)\nmodel.build_vocab(processed_words)\nmodel.train(processed_words,t\notal_examples=model.corpus_count,epochs=20)\nsentence_vectors = []\nfor sentence in processed_words:\n    if len(sentence) !=\n0:\n        sentence_vec = sum([model.wv[word] for word in sentence]) #/ len(sentence)\n        sentence_vectors.append(se\nntence_vec)\n\n#similarity_matrix = [[model.wv.similarity(w1, w2) for w1 in sentence_vectors] for w2 in sentence_vectors]\n#similarity_scores = similarity_matrix.sum(axis=1) # axis = 1 means horizontal\nms = model.wv.most_similar('\\gold\\', topn= 6)\n[1:]\n#top_sentences_indices = np.argsort(ms, axis=0)[-5:] # axis = 0 means sort vertically, and get the top 5\n#top_sentenc\nes = []\nfor index in range(len(ms)):\n    val = processed_sentences[index]\n    summary = '\\'.join(val) + '\\'.\\n    #prin\nt(summary)\n\n#summary = summarize(" ".join(sentences), ratio=0.3, word_count=None, split=False)\n\n'
```

```
In [114]: # DO NOT RUN DOES NOT WORK

"""#summary = summarize_text(processed_sentences, processed_words)
token_words = []
for sent in processed_sentences:
    token_words.append(sent)
summary = summarize_text(processed_sentences, token_words)
print(summary)

"""
```

```
Out[114]: '#summary = summarize_text(processed_sentences, processed_words)\ntoken_words = []\nfor sent in processed_sentences:\n    to\nken_words.append(sent)\nsummary = summarize_text(processed_sentences, token_words)\nprint(summary)\n'
```

```
In [115]: # ms = model.wv.most_similar(user_input, topn= 6)[1:]
```

```
In [ ]:
```