2b-Text-Representation

November 13, 2024

[1]:
```
import numpy as np
import pandas as pd
from textblob import TextBlob
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import TfidfTransformer

pd.options.display.max_columns = 100

import nltk
# nltk.download('omw-1.4')
nltk.download('punkt_tab')
# nltk.download('averaged_perceptron_tagger_eng')
```

```
[nltk_data] Downloading package punkt_tab to /root/nltk_data…
[nltk_data]   Package punkt_tab is already up-to-date!
```

[1]:
```
True
```

[2]:
```
%%capture
!python -m textblob.download_corpora
```

[3]:
```
sentence_1 = 'Jen is a good student.'
sentence_2 = 'Jen is also a great guitarist.'
sentence_3 = 'Good students can sometimes be good guitarists'
```

1 Data Cleaning

We want to singularize guitarists and students.

[4]:
```
sentence_3_tb = TextBlob(sentence_3) # Make a textblob so that we can␣
 ↪singularize the word
sentence_3_singular = [x.singularize() for x in sentence_3_tb.words] #␣
 ↪Singularize each word in the text
sentence_3_clean = ' '.join(sentence_3_singular) # Join it together into a␣
 ↪single string
sentence_3_clean
```

[4]:
```
'Good student can sometime be good guitarist'
```

1.1 Bag of Words Using CountVectorizer

[5]:
```
# Perform the count transformation
vectorizer = CountVectorizer(stop_words='english')
bow_vec = vectorizer.fit_transform([sentence_1, sentence_2, sentence_3_clean])
bow_vec
```

[5]:
```
<Compressed Sparse Row sparse matrix of dtype 'int64'
        with 9 stored elements and shape (3, 5)>
```

[6]:
```
bow_vec.toarray()
```

[6]:
```
array([[1, 0, 0, 1, 1],
       [0, 1, 1, 1, 0],
       [2, 0, 1, 0, 1]])
```

[7]:
```
# Print out results in a data frame
sent_df = pd.DataFrame(bow_vec.toarray(), columns = vectorizer.
 ↪get_feature_names_out())
sent_df
```

[7]:

	good	great	guitarist	jen	student
0	1	0	0	1	1
1	0	1	1	1	0
2	2	0	1	0	1

1.1.1 Your Turn

1. Write 4 sentences of your choice.
2. Run the `CountVectorizer` on your sentences.
3. Print the results in a data frame.

[8]:
```
# Solution 1
my_sents = [
    "It was the best of times.",
    "Call me Ishmael.",
    "To be, or not to be.  That is the question.",
    "We do not choose these things because they are easy."
]
my_sents
```

[8]:
```
['It was the best of times.',
 'Call me Ishmael.',
 'To be, or not to be.  That is the question.',
 'We do not choose these things because they are easy.']
```

[9]:
```
# Solution 2
my_vectorizer = CountVectorizer(stop_words='english')
my_bow_vec = my_vectorizer.fit_transform( my_sents )
my_bow_vec.toarray()
```

[9]:
```
array([[1, 0, 0, 0, 0, 0, 1],
       [0, 0, 0, 1, 0, 0, 0],
       [0, 0, 0, 0, 1, 0, 0],
       [0, 1, 1, 0, 0, 1, 0]])
```

[10]:
```
# Solution 3
my_sent_df = pd.DataFrame(my_bow_vec.toarray(), columns = my_vectorizer.
 ↪get_feature_names_out())
my_sent_df
```

[10]:

	best	choose	easy	ishmael	question	things	times
0	1	0	0	0	0	0	1
1	0	0	0	1	0	0	0
2	0	0	0	0	1	0	0
3	0	1	1	0	0	1	0

1.2 TF-IDF

[11]:
```
# Perform the TF-IDF transformation - Option 1 (TfidfVectorizer)
tf_idf_vec = TfidfVectorizer(stop_words = 'english')
tf_idf_jen = tf_idf_vec.fit_transform([sentence_1, sentence_2,␣
 ↪sentence_3_clean])
tf_idf_jen
```

[11]:
```
<Compressed Sparse Row sparse matrix of dtype 'float64'
        with 9 stored elements and shape (3, 5)>
```

[12]:
```
print(sentence_1)
print(sentence_2)
print(sentence_3_clean)
```

```
Jen is a good student.
Jen is also a great guitarist.
Good student can sometime be good guitarist
```

[13]:
```
# Print out results in a dataframe
tf_df = pd.DataFrame(tf_idf_jen.toarray(), columns = tf_idf_vec.
 ↪get_feature_names_out())
tf_df.shape
```

[13]:
```
(3, 5)
```

[14]:
```
tf_df
```

[14]:

	good	great	guitarist	jen	student
0	0.577350	0.000000	0.000000	0.577350	0.577350
1	0.000000	0.680919	0.517856	0.517856	0.000000
2	0.816497	0.000000	0.408248	0.000000	0.408248

[15]:
```
# Perform the TF-IDF transformation - Option 2 (CountVectorizer + TfidfTransformer - better for large datasets)
tf_idf_tran = TfidfTransformer()
tf_idf_jen = tf_idf_tran.fit_transform(bow_vec)
tf_idf_jen
```

[15]:
```
<Compressed Sparse Row sparse matrix of dtype 'float64'
        with 9 stored elements and shape (3, 5)>
```

[16]:
```
# Print out results in a dataframe
tf_df = pd.DataFrame(tf_idf_jen.toarray(), columns = vectorizer.get_feature_names_out())
tf_df
```

[16]:

	good	great	guitarist	jen	student
0	0.577350	0.000000	0.000000	0.577350	0.577350
1	0.000000	0.680919	0.517856	0.517856	0.000000
2	0.816497	0.000000	0.408248	0.000000	0.408248

[17]:
```
# Get a data frame with the TF-IDF values sorted for document 0
df = pd.DataFrame(tf_idf_jen[0].T.todense(), index=tf_idf_vec.get_feature_names_out(), columns=["TF-IDF"])
df = df.sort_values('TF-IDF', ascending=False)
df
```

[17]:

	TF-IDF
good	0.57735
student	0.57735
jen	0.57735
guitarist	0.00000
great	0.00000

[18]:
```
tf_df.transpose()[0].sort_values(ascending = False)
```

[18]:
```
good         0.57735
student      0.57735
jen          0.57735
guitarist    0.00000
great        0.00000
Name: 0, dtype: float64
```

1.2.1 Your Turn

1. Use the `TfidfTransformer` to transform the bag of words matrix you created above to TF-IDF.
2. Print out the results in a data frame.

[19]:
```
# Solution 1
my_tf_idf_tran = TfidfTransformer()
my_tf_idf_jen = my_tf_idf_tran.fit_transform(my_bow_vec)
my_tf_idf_jen
```

[19]:
```
<Compressed Sparse Row sparse matrix of dtype 'float64'
        with 7 stored elements and shape (4, 7)>
```

[20]:
```
my_tf_idf_jen.toarray()
```

[20]:
```
array([[0.70710678, 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.70710678],
       [0.        , 0.        , 0.        , 1.        , 0.        ,
        0.        , 0.        ],
       [0.        , 0.        , 0.        , 0.        , 1.        ,
        0.        , 0.        ],
       [0.        , 0.57735027, 0.57735027, 0.        , 0.        ,
        0.57735027, 0.        ]])
```

[21]:
```
# Solution 2
# Print out results in a dataframe
my_tf_df = pd.DataFrame(my_tf_idf_jen.toarray(), columns = my_vectorizer.
 ↪get_feature_names_out())
my_tf_df
```

[21]:

	best	choose	easy	ishmael	question	things	times
0	0.707107	0.00000	0.00000	0.0	0.0	0.00000	0.707107
1	0.000000	0.00000	0.00000	1.0	0.0	0.00000	0.000000
2	0.000000	0.00000	0.00000	0.0	1.0	0.00000	0.000000
3	0.000000	0.57735	0.57735	0.0	0.0	0.57735	0.000000

2 Another Example - Using Wikipedia API

[22]:
```
%%capture output
#install Wikipedia API
!pip3 install wikipedia-api
```

[23]:
```
import wikipediaapi
```

[24]:
```
# Pull out the popcorn page from wikipedia - https://en.wikipedia.org/wiki/
 ↪Popcorn
topic = 'popcorn'
```

```
wikip = wikipediaapi.Wikipedia(user_agent = 'foobar')
page_ex = wikip.page(topic)
wiki_text = page_ex.text
wiki_text
```

[24]: 'Popcorn (also called popped corn, popcorns, or pop-corn) is a variety of corn kernel which expands and puffs up when heated. The term also refers to the snack food produced by the expansion. It is one of the oldest snacks, with evidence of popcorn dating back thousands of years in the Americas. It is commonly eaten salted, sweetened, or with artificial flavorings. \nA popcorn kernel\'s strong hull contains the seed\'s hard, starchy shell endosperm with 14-20% moisture, which turns to steam as the kernel is heated. Pressure from the steam continues to build until the hull ruptures, allowing the kernel to forcefully expand, to 20 to 50 times its original size, and then cool.\nSome strains of corn (taxonomized as Zea mays) are cultivated specifically as popping corns. The Zea mays variety everta, a special kind of flint corn, is the most common of these. Popcorn is one of six major types of corn, which includes dent corn, flint corn, pod corn, flour corn, and sweet corn.\n\nHistory\nCorn was domesticated about 10,000 years ago, in what is now Mexico. Archaeologists discovered that people have known about popcorn for thousands of years. Fossil evidence from Peru suggests that corn was present there as early as 4700 BCE, and popped there over 1000 years ago. Between 2007 and 2011, evidence, as early as 4700 BCE, for popping corn, as macrofossil cobs, were discovered at the Paredones and Huaca Prieta archaeological sites on the northern coast of Peru.\nIn 1948 and 1950, evidence, as early as 3600 BCE, for popping corn, as ears of popcorn, were discovered by Harvard anthropology graduate student Herbert W. Dick and Harvard botany graduate student Claude Earle Smith, Junior (1922-1987), in a complex of rock shelters, dubbed the "Bat Cave", in Catron County, west-central New Mexico, and attributed to the Ancestral Puebloan peoples, who maintained trade networks with peoples in tropical Mexico.\nThrough the 19th century, popping of the kernels was achieved by hand, on stove tops over flame. Kernels were sold on the East Coast of the United States under names such as Pearls or Nonpareil. The term popped corn first appeared in John Russell Bartlett\'s 1848 Dictionary of Americanisms. Popcorn is an ingredient in Cracker Jack and, in the early years of the product, it was popped by hand.\n\nPopcorn\'s accessibility increased rapidly in the 1890s with Charles Cretors\' invention of the popcorn maker. Cretors, a Chicago candy store owner, had created a number of steam-powered machines for roasting nuts and applied the technology to the corn kernels.\n\nBy the turn of the century, Cretors had created and deployed street carts equipped with steam-powered popcorn makers.During the Great Depression, popcorn was fairly inexpensive at 5-10 cents a bag and became popular. Thus, while other businesses failed, the popcorn business thrived and became a source of income for many struggling farmers and entrepreneurs, including the Redenbacher family, namesake of the Orville Redenbacher\'s popcorn brand. During World War II, sugar rations diminished candy production, and Americans compensated by eating three times as much popcorn as they had before. The snack was popular at theaters, much to the initial displeasure of many of the theater owners, who thought it

distracted from the films. Their minds eventually changed, however, and in 1938 a Midwestern theater owner named Glen W. Dickinson Sr. installed popcorn machines in the lobbies of his Dickinson theaters. Popcorn was more profitable than theater tickets, and at the suggestion of his production consultant, R. Ray Aden, Dickinson purchased popcorn farms and was able to keep ticket prices down. The venture was a success, and popcorn soon spread. The rise of television in the 1940s brought lower popcorn consumption as theater attendance fell. The Popcorn Institute (a trade association of popcorn processors) promoted popcorn consumption at home, bringing it back to previous levels.\n In 1970, Orville Redenbacher\'s namesake brand of popcorn was launched. In 1981, General Mills received the first patent for a microwave oven popcorn bag; popcorn consumption saw an increase.\nAt least six localities (all in the Midwestern United States) claim to be the "Popcorn Capital of the World;": Ridgway, Illinois; Valparaiso, Indiana; Van Buren, Indiana; Schaller, Iowa; Marion, Ohio; and North Loup, Nebraska. According to the USDA, specific corn for popcorn is grown mostly in Nebraska and Indiana, and increasingly in Texas. As the result of an elementary school project, popcorn became the official state snack food of Illinois.\n\nPopping mechanism\nEach kernel of popcorn contains moisture and oil. Unlike most other grains, the outer hull of the popcorn kernel is strong and impervious to moisture, and the starch inside consists almost entirely of a hard type.\n\nAs the oil and water in the kernel are heated, they turn into steam. Under these conditions, the starch inside the kernel gelatinizes and softens. The steam pressure increases until the breaking point of the hull is reached; a pressure of approximately 930 kPa (135 psi) and a temperature of 180 °C (356 °F). The hull ruptures, causing a sudden drop in pressure inside the kernel and a corresponding rapid expansion of the steam, which expands the starch and proteins of the endosperm into airy foam. As the foam rapidly cools, the starch and protein polymers set into the familiar crispy puff.\nSpecial varieties are grown to improve popping yield. Though the kernels of some other types will pop, the cultivated strain for popcorn is Zea mays everta, which is a variety of flint corn.\n\nCooking methods\nPopcorn can be cooked with butter or oil. Although small quantities can be popped in a stove-top kettle or pot in a home kitchen, commercial sale employs specially designed popcorn machines, which were invented in Chicago, Illinois, by Charles Cretors in 1885. Cretors introduced his invention at the Columbian Exposition in 1893. At that fair, F. W. Rueckheim introduced a molasses-flavored "Candied Popcorn", the first caramel corn; his brother, Louis Ruekheim, slightly altered the recipe and introduced it as Cracker Jack in 1896.\n\nCretors\'s invention was the first patented steam-driven machine that popped corn in oil. Previously, vendors popped corn by holding a wire basket over an open flame. At best, the result was hot, dry, and unevenly cooked. Cretors\'s machine popped corn in a mixture of one-third clarified butter, two-thirds lard, and salt. This mixture can withstand the 232 °C (450 °F) temperature needed to pop corn and produces little smoke. A fire under a boiler created steam that drove a small engine to drive gears, shaft, and the agitator that stirred the corn, and also powered a small puppet, "The Toasty Roasty Man", an attention-getting amusement to attract business. A wire connected to the top of the cooking pan allowed the operator to disengage the

drive mechanism, lift the cover, and dump popped corn into the storage bin beneath. Exhaust from the steam engine was piped to a pan below the corn storage bin and kept freshly popped corn warm. Excess steam was also used to operate a small, shrill whistle to attract attention.\nA different method of popcorn-making involves the "popcorn hammer", a large cast-iron canister that is sealed with a heavy lid and slowly turned over a fire in rotisserie fashion.\n\nExpansion and yield\nPopping results are sensitive to the rate at which the kernels are heated. If heated too quickly, the steam in the outer layers of the kernel can reach high pressures and rupture the hull before the starch in the center of the kernel can fully gelatinize, leading to partially popped kernels with hard centers. Heating too slowly leads to entirely unpopped kernels: the tip of the kernel, where it attached to the cob, is not entirely moisture-proof, and when heated slowly, the steam can leak out of the tip fast enough to keep the pressure from rising sufficiently to break the hull and cause the pop.\nProducers and sellers of popcorn consider two major factors in evaluating the quality of popcorn: what percentage of the kernels will pop, and how much each popped kernel expands. Expansion is an important factor to both the consumer and vendor. For the consumer, larger pieces of popcorn tend to be more tender and are associated with higher quality. For the grower, distributor and vendor, expansion is closely correlated with profit: vendors such as theaters buy popcorn by weight and sell it by volume. For these reasons, higher-expansion popcorn fetches a higher profit per unit weight.\nPopcorn will pop when freshly harvested, but not well; its high moisture content leads to poor expansion and chewy pieces of popcorn. Kernels with a high moisture content are also susceptible to mold when stored. For these reasons, popcorn growers and distributors dry the kernels until they reach the moisture level at which they expand the most. This differs by variety and conditions, but is generally in the range of 14-15% moisture by weight. If the kernels are over-dried, the expansion rate will suffer and the percentage of kernels that pop will decline. Old popcorn tends to dry out, lowering the yield.\nWhen the popcorn has finished popping, sometimes unpopped kernels remain. Known in the popcorn industry as "old maids", these kernels fail to pop because they do not have enough moisture to create enough steam for an explosion. Re-hydrating prior to popping usually results in eliminating the unpopped kernels.\nPopcorn varieties are broadly categorized by the shape of the kernels, the color of the kernels, or the shape of the popped corn. While the kernels may come in a variety of colors, the popped corn is always off-yellow or white as it is only the hull (or pericarp) that is colored. "Rice" type popcorn have a long kernel pointed at both ends; "pearl" type kernels are rounded at the top. Commercial popcorn production has moved mostly to pearl types. Historically, pearl popcorn were usually yellow and rice popcorn usually white. Today both shapes are available in both colors, as well as others including black, red, mauve, purple, and variegated. Mauve and purple popcorn usually have smaller and nutty kernels. Commercial production is dominated by white and yellow.\n\nTerminology\nIn the popcorn industry, a popped kernel of corn is known as a "flake". Two shapes of flakes are commercially important. "Butterfly" (or "snowflake") flakes are irregular in shape and have a number of protruding "wings". "Mushroom" flakes are largely ball-shaped, with

few wings. Butterfly flakes are regarded as having better mouthfeel, with greater tenderness and less noticeable hulls. Mushroom flakes are less fragile than butterfly flakes and are therefore often used for packaged popcorn or confectionery, such as caramel corn. The kernels from a single cob of popcorn may form both butterfly and mushroom flakes; hybrids that produce 100% butterfly flakes or 100% mushroom flakes exist, the latter developed only as recently as 1998.\n\nConsumption\nPopcorn is a popular snack food at sporting events and in movie theaters, where it has been served since the 1930s. Cinemas have come under fire due to their high markup on popcorn; Stuart Hanson, a film historian at De Montfort University in Leicester, once said, "One of the great jokes in the industry is that popcorn is second only to cocaine or heroin in terms of profit."\nTraditions differ as to whether popcorn is consumed as a hearty snack food with salt (predominating in the United States) or as a sweet snack food with caramelized sugar (predominating in Germany).\n\nPopcorn smell has an unusually attractive quality for human beings. This is largely because it contains high levels of the chemicals 6-acetyl-2,3,4,5-tetrahydropyridine and 2-acetyl-1-pyrroline, very powerful aroma compounds that are also used by food and other industries either to make products that smell like popcorn, bread, or other foods containing the compound in nature, or for other purposes.\n\nPopcorn as a breakfast cereal was consumed by Americans in the 1800s and generally consisted of popcorn with milk and a sweetener.\n\nPopcorn balls (popped kernels stuck together with a sugary "glue") were hugely popular around the turn of the 20th century, but their popularity has since waned. Popcorn balls are still served in some places as a traditional Halloween treat. Cracker Jack is a popular, commercially produced candy that consists of peanuts mixed in with caramel-covered popcorn. Kettle corn is a variation of normal popcorn, cooked with white sugar and salt, traditionally in a large copper kettle. Once reserved for specialty shops and county fairs, kettle corn has recently become popular, especially in the microwave popcorn market. The popcorn maker is a relatively new home appliance, and its popularity is increasing because it offers the opportunity to add flavors of the consumer\'s own choice and to choose healthy-eating popcorn styles.\nPopped sorghum is popular as a snack in India. The popped sorghum is similar to popcorn, but the puffs are smaller. Recipes for popping sorghum by microwave, in a pot, etc., are readily available online.\n\nNutritional value\nAir-popped popcorn (no salt or other additives) is 4% water, 78% carbohydrates (including 15% dietary fiber), 12% protein, and 4% fat (table). In a 100 gram reference amount, popcorn provides 382 calories and is a rich source (20% or more of the Daily Value, DV) of riboflavin (25% DV) and several dietary minerals, particularly manganese, phosphorus, and zinc (36-45% DV). B vitamins and other minerals are in appreciable amounts (table).\n\nSaturated fat\nMovie theaters commonly use coconut oil to pop the corn, and then top it with butter or margarine. Movie theater popcorn contains large amounts of saturated fats and sodium due to its method of preparation.\n\nPhytochemicals\nSorghum grains can be popped to form popcorn. All sorghums contain phenolic acids, and most contain flavonoids. Sorghum grains are one of the highest food sources of the flavonoid proanthocyanidin.\n\nHealth risks\nPopcorn is included on the list of foods that the American Academy of

Pediatrics recommends not serving to children under four, because of the risk of choking.\nMicrowaveable popcorn represents a special case, since it is designed to be cooked along with its various flavoring agents. One of these formerly common artificial-butter flavorants, diacetyl, has been implicated in causing respiratory illnesses in microwave popcorn factory workers, also known as "popcorn lung". Major manufacturers in the United States have stopped using this chemical, including Orville Redenbacher\'s, Act II, Pop Secret and Jolly Time.\n\nOther uses\nPopcorn, threaded onto a string, is used as a wall or Christmas tree decoration in some parts of North America, as well as on the Balkan peninsula.\nSome shipping companies have experimented with using popcorn as a biodegradable replacement for expanded polystyrene packing material. However, popcorn has numerous undesirable properties as a packing material, including attractiveness to pests, flammability, and a higher cost and greater density than expanded polystyrene. A more processed form of expanded corn foam has been developed to overcome some of these limitations, forming starch-based foam peanuts.\n\nSee also\nReferences\nFurther reading\n\nHallauer, Arnel R. (2001). Specialty Corns. CRC Press. ISBN 978-0-8493-2377-5.\nLusas, Edmund W.; Rooney, Lloyd W. (2001). Snack Foods Processing. CRC Press. ISBN 978-1-56676-932-7.\nSmith, Andrew F. (1999). Popped Culture: The Social History of Popcorn in America. University of South Carolina Press. ISBN 978-1-57003-300-1.'

2.0.1 Clean the text - version 1

Using string replace.

[25]:
```
# Replace newline chars with spaces before doing any processing. Strip the '␣
 ↪and "s" from possessives
wiki_text_clean = (
    wiki_text
    .replace("\n"," ")
    .replace("'s",'')
    .replace("'",'')
)
wiki_text_clean
```

[25]: 'Popcorn (also called popped corn, popcorns, or pop-corn) is a variety of corn kernel which expands and puffs up when heated. The term also refers to the snack food produced by the expansion. It is one of the oldest snacks, with evidence of popcorn dating back thousands of years in the Americas. It is commonly eaten salted, sweetened, or with artificial flavorings. A popcorn kernel strong hull contains the seed hard, starchy shell endosperm with 14-20% moisture, which turns to steam as the kernel is heated. Pressure from the steam continues to build until the hull ruptures, allowing the kernel to forcefully expand, to 20 to 50 times its original size, and then cool. Some strains of corn (taxonomized as Zea mays) are cultivated specifically as popping corns. The Zea mays variety everta, a special kind of flint corn, is the most common of these. Popcorn is

one of six major types of corn, which includes dent corn, flint corn, pod corn, flour corn, and sweet corn. History Corn was domesticated about 10,000 years ago, in what is now Mexico. Archaeologists discovered that people have known about popcorn for thousands of years. Fossil evidence from Peru suggests that corn was present there as early as 4700 BCE, and popped there over 1000 years ago. Between 2007 and 2011, evidence, as early as 4700 BCE, for popping corn, as macrofossil cobs, were discovered at the Paredones and Huaca Prieta archaeological sites on the northern coast of Peru. In 1948 and 1950, evidence, as early as 3600 BCE, for popping corn, as ears of popcorn, were discovered by Harvard anthropology graduate student Herbert W. Dick and Harvard botany graduate student Claude Earle Smith, Junior (1922-1987), in a complex of rock shelters, dubbed the "Bat Cave", in Catron County, west-central New Mexico, and attributed to the Ancestral Puebloan peoples, who maintained trade networks with peoples in tropical Mexico. Through the 19th century, popping of the kernels was achieved by hand, on stove tops over flame. Kernels were sold on the East Coast of the United States under names such as Pearls or Nonpareil. The term popped corn first appeared in John Russell Bartlett 1848 Dictionary of Americanisms. Popcorn is an ingredient in Cracker Jack and, in the early years of the product, it was popped by hand. Popcorn accessibility increased rapidly in the 1890s with Charles Cretors invention of the popcorn maker. Cretors, a Chicago candy store owner, had created a number of steam-powered machines for roasting nuts and applied the technology to the corn kernels. By the turn of the century, Cretors had created and deployed street carts equipped with steam-powered popcorn makers.During the Great Depression, popcorn was fairly inexpensive at 5-10 cents a bag and became popular. Thus, while other businesses failed, the popcorn business thrived and became a source of income for many struggling farmers and entrepreneurs, including the Redenbacher family, namesake of the Orville Redenbacher popcorn brand. During World War II, sugar rations diminished candy production, and Americans compensated by eating three times as much popcorn as they had before. The snack was popular at theaters, much to the initial displeasure of many of the theater owners, who thought it distracted from the films. Their minds eventually changed, however, and in 1938 a Midwestern theater owner named Glen W. Dickinson Sr. installed popcorn machines in the lobbies of his Dickinson theaters. Popcorn was more profitable than theater tickets, and at the suggestion of his production consultant, R. Ray Aden, Dickinson purchased popcorn farms and was able to keep ticket prices down. The venture was a success, and popcorn soon spread. The rise of television in the 1940s brought lower popcorn consumption as theater attendance fell. The Popcorn Institute (a trade association of popcorn processors) promoted popcorn consumption at home, bringing it back to previous levels. In 1970, Orville Redenbacher namesake brand of popcorn was launched. In 1981, General Mills received the first patent for a microwave oven popcorn bag; popcorn consumption saw an increase. At least six localities (all in the Midwestern United States) claim to be the "Popcorn Capital of the World;": Ridgway, Illinois; Valparaiso, Indiana; Van Buren, Indiana; Schaller, Iowa; Marion, Ohio; and North Loup, Nebraska. According to the USDA, specific corn for popcorn is grown mostly in Nebraska and Indiana, and increasingly in Texas. As the result of an elementary

school project, popcorn became the official state snack food of Illinois. Popping mechanism Each kernel of popcorn contains moisture and oil. Unlike most other grains, the outer hull of the popcorn kernel is strong and impervious to moisture, and the starch inside consists almost entirely of a hard type. As the oil and water in the kernel are heated, they turn into steam. Under these conditions, the starch inside the kernel gelatinizes and softens. The steam pressure increases until the breaking point of the hull is reached; a pressure of approximately 930 kPa (135 psi) and a temperature of 180 °C (356 °F). The hull ruptures, causing a sudden drop in pressure inside the kernel and a corresponding rapid expansion of the steam, which expands the starch and proteins of the endosperm into airy foam. As the foam rapidly cools, the starch and protein polymers set into the familiar crispy puff. Special varieties are grown to improve popping yield. Though the kernels of some other types will pop, the cultivated strain for popcorn is Zea mays everta, which is a variety of flint corn. Cooking methods Popcorn can be cooked with butter or oil. Although small quantities can be popped in a stove-top kettle or pot in a home kitchen, commercial sale employs specially designed popcorn machines, which were invented in Chicago, Illinois, by Charles Cretors in 1885. Cretors introduced his invention at the Columbian Exposition in 1893. At that fair, F. W. Rueckheim introduced a molasses-flavored "Candied Popcorn", the first caramel corn; his brother, Louis Ruekheim, slightly altered the recipe and introduced it as Cracker Jack in 1896. Cretors invention was the first patented steam-driven machine that popped corn in oil. Previously, vendors popped corn by holding a wire basket over an open flame. At best, the result was hot, dry, and unevenly cooked. Cretors machine popped corn in a mixture of one-third clarified butter, two-thirds lard, and salt. This mixture can withstand the 232 °C (450 °F) temperature needed to pop corn and produces little smoke. A fire under a boiler created steam that drove a small engine to drive gears, shaft, and the agitator that stirred the corn, and also powered a small puppet, "The Toasty Roasty Man", an attention-getting amusement to attract business. A wire connected to the top of the cooking pan allowed the operator to disengage the drive mechanism, lift the cover, and dump popped corn into the storage bin beneath. Exhaust from the steam engine was piped to a pan below the corn storage bin and kept freshly popped corn warm. Excess steam was also used to operate a small, shrill whistle to attract attention. A different method of popcorn-making involves the "popcorn hammer", a large cast-iron canister that is sealed with a heavy lid and slowly turned over a fire in rotisserie fashion. Expansion and yield Popping results are sensitive to the rate at which the kernels are heated. If heated too quickly, the steam in the outer layers of the kernel can reach high pressures and rupture the hull before the starch in the center of the kernel can fully gelatinize, leading to partially popped kernels with hard centers. Heating too slowly leads to entirely unpopped kernels: the tip of the kernel, where it attached to the cob, is not entirely moisture-proof, and when heated slowly, the steam can leak out of the tip fast enough to keep the pressure from rising sufficiently to break the hull and cause the pop. Producers and sellers of popcorn consider two major factors in evaluating the quality of popcorn: what percentage of the kernels will pop, and how much each popped kernel expands.

Expansion is an important factor to both the consumer and vendor. For the consumer, larger pieces of popcorn tend to be more tender and are associated with higher quality. For the grower, distributor and vendor, expansion is closely correlated with profit: vendors such as theaters buy popcorn by weight and sell it by volume. For these reasons, higher-expansion popcorn fetches a higher profit per unit weight. Popcorn will pop when freshly harvested, but not well; its high moisture content leads to poor expansion and chewy pieces of popcorn. Kernels with a high moisture content are also susceptible to mold when stored. For these reasons, popcorn growers and distributors dry the kernels until they reach the moisture level at which they expand the most. This differs by variety and conditions, but is generally in the range of 14-15% moisture by weight. If the kernels are over-dried, the expansion rate will suffer and the percentage of kernels that pop will decline. Old popcorn tends to dry out, lowering the yield. When the popcorn has finished popping, sometimes unpopped kernels remain. Known in the popcorn industry as "old maids", these kernels fail to pop because they do not have enough moisture to create enough steam for an explosion. Re-hydrating prior to popping usually results in eliminating the unpopped kernels. Popcorn varieties are broadly categorized by the shape of the kernels, the color of the kernels, or the shape of the popped corn. While the kernels may come in a variety of colors, the popped corn is always off-yellow or white as it is only the hull (or pericarp) that is colored. "Rice" type popcorn have a long kernel pointed at both ends; "pearl" type kernels are rounded at the top. Commercial popcorn production has moved mostly to pearl types. Historically, pearl popcorn were usually yellow and rice popcorn usually white. Today both shapes are available in both colors, as well as others including black, red, mauve, purple, and variegated. Mauve and purple popcorn usually have smaller and nutty kernels. Commercial production is dominated by white and yellow. Terminology In the popcorn industry, a popped kernel of corn is known as a "flake". Two shapes of flakes are commercially important. "Butterfly" (or "snowflake") flakes are irregular in shape and have a number of protruding "wings". "Mushroom" flakes are largely ball-shaped, with few wings. Butterfly flakes are regarded as having better mouthfeel, with greater tenderness and less noticeable hulls. Mushroom flakes are less fragile than butterfly flakes and are therefore often used for packaged popcorn or confectionery, such as caramel corn. The kernels from a single cob of popcorn may form both butterfly and mushroom flakes; hybrids that produce 100% butterfly flakes or 100% mushroom flakes exist, the latter developed only as recently as 1998. Consumption Popcorn is a popular snack food at sporting events and in movie theaters, where it has been served since the 1930s. Cinemas have come under fire due to their high markup on popcorn; Stuart Hanson, a film historian at De Montfort University in Leicester, once said, "One of the great jokes in the industry is that popcorn is second only to cocaine or heroin in terms of profit." Traditions differ as to whether popcorn is consumed as a hearty snack food with salt (predominating in the United States) or as a sweet snack food with caramelized sugar (predominating in Germany). Popcorn smell has an unusually attractive quality for human beings. This is largely because it contains high levels of the chemicals 6-acetyl-2,3,4,5-tetrahydropyridine and 2-acetyl-1-pyrroline, very

powerful aroma compounds that are also used by food and other industries either to make products that smell like popcorn, bread, or other foods containing the compound in nature, or for other purposes. Popcorn as a breakfast cereal was consumed by Americans in the 1800s and generally consisted of popcorn with milk and a sweetener. Popcorn balls (popped kernels stuck together with a sugary "glue") were hugely popular around the turn of the 20th century, but their popularity has since waned. Popcorn balls are still served in some places as a traditional Halloween treat. Cracker Jack is a popular, commercially produced candy that consists of peanuts mixed in with caramel-covered popcorn. Kettle corn is a variation of normal popcorn, cooked with white sugar and salt, traditionally in a large copper kettle. Once reserved for specialty shops and county fairs, kettle corn has recently become popular, especially in the microwave popcorn market. The popcorn maker is a relatively new home appliance, and its popularity is increasing because it offers the opportunity to add flavors of the consumer own choice and to choose healthy-eating popcorn styles. Popped sorghum is popular as a snack in India. The popped sorghum is similar to popcorn, but the puffs are smaller. Recipes for popping sorghum by microwave, in a pot, etc., are readily available online. Nutritional value Air-popped popcorn (no salt or other additives) is 4% water, 78% carbohydrates (including 15% dietary fiber), 12% protein, and 4% fat (table). In a 100 gram reference amount, popcorn provides 382 calories and is a rich source (20% or more of the Daily Value, DV) of riboflavin (25% DV) and several dietary minerals, particularly manganese, phosphorus, and zinc (36-45% DV). B vitamins and other minerals are in appreciable amounts (table). Saturated fat Movie theaters commonly use coconut oil to pop the corn, and then top it with butter or margarine. Movie theater popcorn contains large amounts of saturated fats and sodium due to its method of preparation. Phytochemicals Sorghum grains can be popped to form popcorn. All sorghums contain phenolic acids, and most contain flavonoids. Sorghum grains are one of the highest food sources of the flavonoid proanthocyanidin. Health risks Popcorn is included on the list of foods that the American Academy of Pediatrics recommends not serving to children under four, because of the risk of choking. Microwaveable popcorn represents a special case, since it is designed to be cooked along with its various flavoring agents. One of these formerly common artificial-butter flavorants, diacetyl, has been implicated in causing respiratory illnesses in microwave popcorn factory workers, also known as "popcorn lung". Major manufacturers in the United States have stopped using this chemical, including Orville Redenbacher, Act II, Pop Secret and Jolly Time. Other uses Popcorn, threaded onto a string, is used as a wall or Christmas tree decoration in some parts of North America, as well as on the Balkan peninsula. Some shipping companies have experimented with using popcorn as a biodegradable replacement for expanded polystyrene packing material. However, popcorn has numerous undesirable properties as a packing material, including attractiveness to pests, flammability, and a higher cost and greater density than expanded polystyrene. A more processed form of expanded corn foam has been developed to overcome some of these limitations, forming starch-based foam peanuts. See also References Further reading Hallauer, Arnel R. (2001). Specialty Corns. CRC Press. ISBN 978-0-8493-2377-5. Lusas, Edmund W.;

Rooney, Lloyd W. (2001). Snack Foods Processing. CRC Press. ISBN 978-1-56676-932-7. Smith, Andrew F. (1999). Popped Culture: The Social History of Popcorn in America. University of South Carolina Press. ISBN 978-1-57003-300-1.'

2.0.2 Clean the text - version 2

Using a for..loop and string replace.

[26]:
```
wiki_text_clean = wiki_text.lower()
for c in ["\n", "'s", "'", "  "]:
  wiki_text_clean = wiki_text_clean.replace(c," ")
wiki_text_clean
```

[26]:
```
'popcorn (also called popped corn, popcorns, or pop-corn) is a variety of corn kernel which expands and puffs up when heated. the term also refers to the snack food produced by the expansion. it is one of the oldest snacks, with evidence of popcorn dating back thousands of years in the americas. it is commonly eaten salted, sweetened, or with artificial flavorings. a popcorn kernel strong hull contains the seed hard, starchy shell endosperm with 14-20% moisture, which turns to steam as the kernel is heated. pressure from the steam continues to build until the hull ruptures, allowing the kernel to forcefully expand, to 20 to 50 times its original size, and then cool. some strains of corn (taxonomized as zea mays) are cultivated specifically as popping corns. the zea mays variety everta, a special kind of flint corn, is the most common of these. popcorn is one of six major types of corn, which includes dent corn, flint corn, pod corn, flour corn, and sweet corn. history corn was domesticated about 10,000 years ago, in what is now mexico. archaeologists discovered that people have known about popcorn for thousands of years. fossil evidence from peru suggests that corn was present there as early as 4700 bce, and popped there over 1000 years ago. between 2007 and 2011, evidence, as early as 4700 bce, for popping corn, as macrofossil cobs, were discovered at the paredones and huaca prieta archaeological sites on the northern coast of peru. in 1948 and 1950, evidence, as early as 3600 bce, for popping corn, as ears of popcorn, were discovered by harvard anthropology graduate student herbert w. dick and harvard botany graduate student claude earle smith, junior (1922-1987), in a complex of rock shelters, dubbed the "bat cave", in catron county, west-central new mexico, and attributed to the ancestral puebloan peoples, who maintained trade networks with peoples in tropical mexico. through the 19th century, popping of the kernels was achieved by hand, on stove tops over flame. kernels were sold on the east coast of the united states under names such as pearls or nonpareil. the term popped corn first appeared in john russell bartlett 1848 dictionary of americanisms. popcorn is an ingredient in cracker jack and, in the early years of the product, it was popped by hand. popcorn accessibility increased rapidly in the 1890s with charles cretors invention of the popcorn maker. cretors, a chicago candy store owner, had created a number of steam-powered machines for roasting nuts and applied the technology to the corn kernels. by the turn of the century, cretors had created and deployed street carts equipped with steam-powered popcorn
```

makers.during the great depression, popcorn was fairly inexpensive at 5-10 cents a bag and became popular. thus, while other businesses failed, the popcorn business thrived and became a source of income for many struggling farmers and entrepreneurs, including the redenbacher family, namesake of the orville redenbacher popcorn brand. during world war ii, sugar rations diminished candy production, and americans compensated by eating three times as much popcorn as they had before. the snack was popular at theaters, much to the initial displeasure of many of the theater owners, who thought it distracted from the films. their minds eventually changed, however, and in 1938 a midwestern theater owner named glen w. dickinson sr. installed popcorn machines in the lobbies of his dickinson theaters. popcorn was more profitable than theater tickets, and at the suggestion of his production consultant, r. ray aden, dickinson purchased popcorn farms and was able to keep ticket prices down. the venture was a success, and popcorn soon spread. the rise of television in the 1940s brought lower popcorn consumption as theater attendance fell. the popcorn institute (a trade association of popcorn processors) promoted popcorn consumption at home, bringing it back to previous levels. in 1970, orville redenbacher namesake brand of popcorn was launched. in 1981, general mills received the first patent for a microwave oven popcorn bag; popcorn consumption saw an increase. at least six localities (all in the midwestern united states) claim to be the "popcorn capital of the world;": ridgway, illinois; valparaiso, indiana; van buren, indiana; schaller, iowa; marion, ohio; and north loup, nebraska. according to the usda, specific corn for popcorn is grown mostly in nebraska and indiana, and increasingly in texas. as the result of an elementary school project, popcorn became the official state snack food of illinois. popping mechanism each kernel of popcorn contains moisture and oil. unlike most other grains, the outer hull of the popcorn kernel is strong and impervious to moisture, and the starch inside consists almost entirely of a hard type. as the oil and water in the kernel are heated, they turn into steam. under these conditions, the starch inside the kernel gelatinizes and softens. the steam pressure increases until the breaking point of the hull is reached; a pressure of approximately 930 kpa (135 psi) and a temperature of 180 °c (356 °f). the hull ruptures, causing a sudden drop in pressure inside the kernel and a corresponding rapid expansion of the steam, which expands the starch and proteins of the endosperm into airy foam. as the foam rapidly cools, the starch and protein polymers set into the familiar crispy puff. special varieties are grown to improve popping yield. though the kernels of some other types will pop, the cultivated strain for popcorn is zea mays everta, which is a variety of flint corn. cooking methods popcorn can be cooked with butter or oil. although small quantities can be popped in a stove-top kettle or pot in a home kitchen, commercial sale employs specially designed popcorn machines, which were invented in chicago, illinois, by charles cretors in 1885. cretors introduced his invention at the columbian exposition in 1893. at that fair, f. w. rueckheim introduced a molasses-flavored "candied popcorn", the first caramel corn; his brother, louis ruekheim, slightly altered the recipe and introduced it as cracker jack in 1896. cretors invention was the first patented steam-driven machine that popped corn in oil. previously, vendors popped corn by holding a wire basket over an open flame. at best, the

result was hot, dry, and unevenly cooked. cretors machine popped corn in a mixture of one-third clarified butter, two-thirds lard, and salt. this mixture can withstand the 232 °c (450 °f) temperature needed to pop corn and produces little smoke. a fire under a boiler created steam that drove a small engine to drive gears, shaft, and the agitator that stirred the corn, and also powered a small puppet, "the toasty roasty man", an attention-getting amusement to attract business. a wire connected to the top of the cooking pan allowed the operator to disengage the drive mechanism, lift the cover, and dump popped corn into the storage bin beneath. exhaust from the steam engine was piped to a pan below the corn storage bin and kept freshly popped corn warm. excess steam was also used to operate a small, shrill whistle to attract attention. a different method of popcorn-making involves the "popcorn hammer", a large cast-iron canister that is sealed with a heavy lid and slowly turned over a fire in rotisserie fashion. expansion and yield popping results are sensitive to the rate at which the kernels are heated. if heated too quickly, the steam in the outer layers of the kernel can reach high pressures and rupture the hull before the starch in the center of the kernel can fully gelatinize, leading to partially popped kernels with hard centers. heating too slowly leads to entirely unpopped kernels: the tip of the kernel, where it attached to the cob, is not entirely moisture-proof, and when heated slowly, the steam can leak out of the tip fast enough to keep the pressure from rising sufficiently to break the hull and cause the pop. producers and sellers of popcorn consider two major factors in evaluating the quality of popcorn: what percentage of the kernels will pop, and how much each popped kernel expands. expansion is an important factor to both the consumer and vendor. for the consumer, larger pieces of popcorn tend to be more tender and are associated with higher quality. for the grower, distributor and vendor, expansion is closely correlated with profit: vendors such as theaters buy popcorn by weight and sell it by volume. for these reasons, higher-expansion popcorn fetches a higher profit per unit weight. popcorn will pop when freshly harvested, but not well; its high moisture content leads to poor expansion and chewy pieces of popcorn. kernels with a high moisture content are also susceptible to mold when stored. for these reasons, popcorn growers and distributors dry the kernels until they reach the moisture level at which they expand the most. this differs by variety and conditions, but is generally in the range of 14-15% moisture by weight. if the kernels are over-dried, the expansion rate will suffer and the percentage of kernels that pop will decline. old popcorn tends to dry out, lowering the yield. when the popcorn has finished popping, sometimes unpopped kernels remain. known in the popcorn industry as "old maids", these kernels fail to pop because they do not have enough moisture to create enough steam for an explosion. re-hydrating prior to popping usually results in eliminating the unpopped kernels. popcorn varieties are broadly categorized by the shape of the kernels, the color of the kernels, or the shape of the popped corn. while the kernels may come in a variety of colors, the popped corn is always off-yellow or white as it is only the hull (or pericarp) that is colored. "rice" type popcorn have a long kernel pointed at both ends; "pearl" type kernels are rounded at the top. commercial popcorn production has moved mostly to pearl types. historically, pearl popcorn were usually yellow and

rice popcorn usually white. today both shapes are available in both colors, as well as others including black, red, mauve, purple, and variegated. mauve and purple popcorn usually have smaller and nutty kernels. commercial production is dominated by white and yellow. terminology in the popcorn industry, a popped kernel of corn is known as a "flake". two shapes of flakes are commercially important. "butterfly" (or "snowflake") flakes are irregular in shape and have a number of protruding "wings". "mushroom" flakes are largely ball-shaped, with few wings. butterfly flakes are regarded as having better mouthfeel, with greater tenderness and less noticeable hulls. mushroom flakes are less fragile than butterfly flakes and are therefore often used for packaged popcorn or confectionery, such as caramel corn. the kernels from a single cob of popcorn may form both butterfly and mushroom flakes; hybrids that produce 100% butterfly flakes or 100% mushroom flakes exist, the latter developed only as recently as 1998. consumption popcorn is a popular snack food at sporting events and in movie theaters, where it has been served since the 1930s. cinemas have come under fire due to their high markup on popcorn; stuart hanson, a film historian at de montfort university in leicester, once said, "one of the great jokes in the industry is that popcorn is second only to cocaine or heroin in terms of profit." traditions differ as to whether popcorn is consumed as a hearty snack food with salt (predominating in the united states) or as a sweet snack food with caramelized sugar (predominating in germany). popcorn smell has an unusually attractive quality for human beings. this is largely because it contains high levels of the chemicals 6-acetyl-2,3,4,5-tetrahydropyridine and 2-acetyl-1-pyrroline, very powerful aroma compounds that are also used by food and other industries either to make products that smell like popcorn, bread, or other foods containing the compound in nature, or for other purposes. popcorn as a breakfast cereal was consumed by americans in the 1800s and generally consisted of popcorn with milk and a sweetener. popcorn balls (popped kernels stuck together with a sugary "glue") were hugely popular around the turn of the 20th century, but their popularity has since waned. popcorn balls are still served in some places as a traditional halloween treat. cracker jack is a popular, commercially produced candy that consists of peanuts mixed in with caramel-covered popcorn. kettle corn is a variation of normal popcorn, cooked with white sugar and salt, traditionally in a large copper kettle. once reserved for specialty shops and county fairs, kettle corn has recently become popular, especially in the microwave popcorn market. the popcorn maker is a relatively new home appliance, and its popularity is increasing because it offers the opportunity to add flavors of the consumer own choice and to choose healthy-eating popcorn styles. popped sorghum is popular as a snack in india. the popped sorghum is similar to popcorn, but the puffs are smaller. recipes for popping sorghum by microwave, in a pot, etc., are readily available online. nutritional value air-popped popcorn (no salt or other additives) is 4% water, 78% carbohydrates (including 15% dietary fiber), 12% protein, and 4% fat (table). in a 100 gram reference amount, popcorn provides 382 calories and is a rich source (20% or more of the daily value, dv) of riboflavin (25% dv) and several dietary minerals, particularly manganese, phosphorus, and zinc (36-45% dv). b vitamins and other minerals are in appreciable amounts (table). saturated fat movie

theaters commonly use coconut oil to pop the corn, and then top it with butter or margarine. movie theater popcorn contains large amounts of saturated fats and sodium due to its method of preparation. phytochemicals sorghum grains can be popped to form popcorn. all sorghums contain phenolic acids, and most contain flavonoids. sorghum grains are one of the highest food sources of the flavonoid proanthocyanidin. health risks popcorn is included on the list of foods that the american academy of pediatrics recommends not serving to children under four, because of the risk of choking. microwaveable popcorn represents a special case, since it is designed to be cooked along with its various flavoring agents. one of these formerly common artificial-butter flavorants, diacetyl, has been implicated in causing respiratory illnesses in microwave popcorn factory workers, also known as "popcorn lung". major manufacturers in the united states have stopped using this chemical, including orville redenbacher , act ii, pop secret and jolly time. other uses popcorn, threaded onto a string, is used as a wall or christmas tree decoration in some parts of north america, as well as on the balkan peninsula. some shipping companies have experimented with using popcorn as a biodegradable replacement for expanded polystyrene packing material. however, popcorn has numerous undesirable properties as a packing material, including attractiveness to pests, flammability, and a higher cost and greater density than expanded polystyrene. a more processed form of expanded corn foam has been developed to overcome some of these limitations, forming starch-based foam peanuts. see also references further reading hallauer, arnel r. (2001). specialty corns. crc press. isbn 978-0-8493-2377-5. lusas, edmund w.; rooney, lloyd w. (2001). snack foods processing. crc press. isbn 978-1-56676-932-7. smith, andrew f. (1999). popped culture: the social history of popcorn in america. university of south carolina press. isbn 978-1-57003-300-1.'

2.0.3 Clean the text - version 3

Using a regular expression.

[27]:
```
import re

pat = re.compile(r"(\n|'s|'| )+")
wiki_text_clean = re.sub(pat,' ', wiki_text.lower())
wiki_text_clean
```

[27]: 'popcorn (also called popped corn, popcorns, or pop-corn) is a variety of corn kernel which expands and puffs up when heated. the term also refers to the snack food produced by the expansion. it is one of the oldest snacks, with evidence of popcorn dating back thousands of years in the americas. it is commonly eaten salted, sweetened, or with artificial flavorings. a popcorn kernel strong hull contains the seed hard, starchy shell endosperm with 14-20% moisture, which turns to steam as the kernel is heated. pressure from the steam continues to build until the hull ruptures, allowing the kernel to forcefully expand, to 20 to 50 times its original size, and then cool. some strains of corn (taxonomized as zea mays) are cultivated specifically as popping corns. the zea mays variety

everta, a special kind of flint corn, is the most common of these. popcorn is one of six major types of corn, which includes dent corn, flint corn, pod corn, flour corn, and sweet corn. history corn was domesticated about 10,000 years ago, in what is now mexico. archaeologists discovered that people have known about popcorn for thousands of years. fossil evidence from peru suggests that corn was present there as early as 4700 bce, and popped there over 1000 years ago. between 2007 and 2011, evidence, as early as 4700 bce, for popping corn, as macrofossil cobs, were discovered at the paredones and huaca prieta archaeological sites on the northern coast of peru. in 1948 and 1950, evidence, as early as 3600 bce, for popping corn, as ears of popcorn, were discovered by harvard anthropology graduate student herbert w. dick and harvard botany graduate student claude earle smith, junior (1922-1987), in a complex of rock shelters, dubbed the "bat cave", in catron county, west-central new mexico, and attributed to the ancestral puebloan peoples, who maintained trade networks with peoples in tropical mexico. through the 19th century, popping of the kernels was achieved by hand, on stove tops over flame. kernels were sold on the east coast of the united states under names such as pearls or nonpareil. the term popped corn first appeared in john russell bartlett 1848 dictionary of americanisms. popcorn is an ingredient in cracker jack and, in the early years of the product, it was popped by hand. popcorn accessibility increased rapidly in the 1890s with charles cretors invention of the popcorn maker. cretors, a chicago candy store owner, had created a number of steam-powered machines for roasting nuts and applied the technology to the corn kernels. by the turn of the century, cretors had created and deployed street carts equipped with steam-powered popcorn makers.during the great depression, popcorn was fairly inexpensive at 5-10 cents a bag and became popular. thus, while other businesses failed, the popcorn business thrived and became a source of income for many struggling farmers and entrepreneurs, including the redenbacher family, namesake of the orville redenbacher popcorn brand. during world war ii, sugar rations diminished candy production, and americans compensated by eating three times as much popcorn as they had before. the snack was popular at theaters, much to the initial displeasure of many of the theater owners, who thought it distracted from the films. their minds eventually changed, however, and in 1938 a midwestern theater owner named glen w. dickinson sr. installed popcorn machines in the lobbies of his dickinson theaters. popcorn was more profitable than theater tickets, and at the suggestion of his production consultant, r. ray aden, dickinson purchased popcorn farms and was able to keep ticket prices down. the venture was a success, and popcorn soon spread. the rise of television in the 1940s brought lower popcorn consumption as theater attendance fell. the popcorn institute (a trade association of popcorn processors) promoted popcorn consumption at home, bringing it back to previous levels. in 1970, orville redenbacher namesake brand of popcorn was launched. in 1981, general mills received the first patent for a microwave oven popcorn bag; popcorn consumption saw an increase. at least six localities (all in the midwestern united states) claim to be the "popcorn capital of the world;": ridgway, illinois; valparaiso, indiana; van buren, indiana; schaller, iowa; marion, ohio; and north loup, nebraska. according to the usda, specific corn for popcorn is grown mostly in nebraska and indiana, and

increasingly in texas. as the result of an elementary school project, popcorn became the official state snack food of illinois. popping mechanism each kernel of popcorn contains moisture and oil. unlike most other grains, the outer hull of the popcorn kernel is strong and impervious to moisture, and the starch inside consists almost entirely of a hard type. as the oil and water in the kernel are heated, they turn into steam. under these conditions, the starch inside the kernel gelatinizes and softens. the steam pressure increases until the breaking point of the hull is reached; a pressure of approximately 930 kpa (135 psi) and a temperature of 180 °c (356 °f). the hull ruptures, causing a sudden drop in pressure inside the kernel and a corresponding rapid expansion of the steam, which expands the starch and proteins of the endosperm into airy foam. as the foam rapidly cools, the starch and protein polymers set into the familiar crispy puff. special varieties are grown to improve popping yield. though the kernels of some other types will pop, the cultivated strain for popcorn is zea mays everta, which is a variety of flint corn. cooking methods popcorn can be cooked with butter or oil. although small quantities can be popped in a stove-top kettle or pot in a home kitchen, commercial sale employs specially designed popcorn machines, which were invented in chicago, illinois, by charles cretors in 1885. cretors introduced his invention at the columbian exposition in 1893. at that fair, f. w. rueckheim introduced a molasses-flavored "candied popcorn", the first caramel corn; his brother, louis ruekheim, slightly altered the recipe and introduced it as cracker jack in 1896. cretors invention was the first patented steam-driven machine that popped corn in oil. previously, vendors popped corn by holding a wire basket over an open flame. at best, the result was hot, dry, and unevenly cooked. cretors machine popped corn in a mixture of one-third clarified butter, two-thirds lard, and salt. this mixture can withstand the 232 °c (450 °f) temperature needed to pop corn and produces little smoke. a fire under a boiler created steam that drove a small engine to drive gears, shaft, and the agitator that stirred the corn, and also powered a small puppet, "the toasty roasty man", an attention-getting amusement to attract business. a wire connected to the top of the cooking pan allowed the operator to disengage the drive mechanism, lift the cover, and dump popped corn into the storage bin beneath. exhaust from the steam engine was piped to a pan below the corn storage bin and kept freshly popped corn warm. excess steam was also used to operate a small, shrill whistle to attract attention. a different method of popcorn-making involves the "popcorn hammer", a large cast-iron canister that is sealed with a heavy lid and slowly turned over a fire in rotisserie fashion. expansion and yield popping results are sensitive to the rate at which the kernels are heated. if heated too quickly, the steam in the outer layers of the kernel can reach high pressures and rupture the hull before the starch in the center of the kernel can fully gelatinize, leading to partially popped kernels with hard centers. heating too slowly leads to entirely unpopped kernels: the tip of the kernel, where it attached to the cob, is not entirely moisture-proof, and when heated slowly, the steam can leak out of the tip fast enough to keep the pressure from rising sufficiently to break the hull and cause the pop. producers and sellers of popcorn consider two major factors in evaluating the quality of popcorn: what percentage of the kernels will pop, and how much each

popped kernel expands. expansion is an important factor to both the consumer and vendor. for the consumer, larger pieces of popcorn tend to be more tender and are associated with higher quality. for the grower, distributor and vendor, expansion is closely correlated with profit: vendors such as theaters buy popcorn by weight and sell it by volume. for these reasons, higher-expansion popcorn fetches a higher profit per unit weight. popcorn will pop when freshly harvested, but not well; its high moisture content leads to poor expansion and chewy pieces of popcorn. kernels with a high moisture content are also susceptible to mold when stored. for these reasons, popcorn growers and distributors dry the kernels until they reach the moisture level at which they expand the most. this differs by variety and conditions, but is generally in the range of 14-15% moisture by weight. if the kernels are over-dried, the expansion rate will suffer and the percentage of kernels that pop will decline. old popcorn tends to dry out, lowering the yield. when the popcorn has finished popping, sometimes unpopped kernels remain. known in the popcorn industry as "old maids", these kernels fail to pop because they do not have enough moisture to create enough steam for an explosion. re-hydrating prior to popping usually results in eliminating the unpopped kernels. popcorn varieties are broadly categorized by the shape of the kernels, the color of the kernels, or the shape of the popped corn. while the kernels may come in a variety of colors, the popped corn is always off-yellow or white as it is only the hull (or pericarp) that is colored. "rice" type popcorn have a long kernel pointed at both ends; "pearl" type kernels are rounded at the top. commercial popcorn production has moved mostly to pearl types. historically, pearl popcorn were usually yellow and rice popcorn usually white. today both shapes are available in both colors, as well as others including black, red, mauve, purple, and variegated. mauve and purple popcorn usually have smaller and nutty kernels. commercial production is dominated by white and yellow. terminology in the popcorn industry, a popped kernel of corn is known as a "flake". two shapes of flakes are commercially important. "butterfly" (or "snowflake") flakes are irregular in shape and have a number of protruding "wings". "mushroom" flakes are largely ball-shaped, with few wings. butterfly flakes are regarded as having better mouthfeel, with greater tenderness and less noticeable hulls. mushroom flakes are less fragile than butterfly flakes and are therefore often used for packaged popcorn or confectionery, such as caramel corn. the kernels from a single cob of popcorn may form both butterfly and mushroom flakes; hybrids that produce 100% butterfly flakes or 100% mushroom flakes exist, the latter developed only as recently as 1998. consumption popcorn is a popular snack food at sporting events and in movie theaters, where it has been served since the 1930s. cinemas have come under fire due to their high markup on popcorn; stuart hanson, a film historian at de montfort university in leicester, once said, "one of the great jokes in the industry is that popcorn is second only to cocaine or heroin in terms of profit." traditions differ as to whether popcorn is consumed as a hearty snack food with salt (predominating in the united states) or as a sweet snack food with caramelized sugar (predominating in germany). popcorn smell has an unusually attractive quality for human beings. this is largely because it contains high levels of the chemicals 6-acetyl-2,3,4,5-tetrahydropyridine and

2-acetyl-1-pyrroline, very powerful aroma compounds that are also used by food and other industries either to make products that smell like popcorn, bread, or other foods containing the compound in nature, or for other purposes. popcorn as a breakfast cereal was consumed by americans in the 1800s and generally consisted of popcorn with milk and a sweetener. popcorn balls (popped kernels stuck together with a sugary "glue") were hugely popular around the turn of the 20th century, but their popularity has since waned. popcorn balls are still served in some places as a traditional halloween treat. cracker jack is a popular, commercially produced candy that consists of peanuts mixed in with caramel-covered popcorn. kettle corn is a variation of normal popcorn, cooked with white sugar and salt, traditionally in a large copper kettle. once reserved for specialty shops and county fairs, kettle corn has recently become popular, especially in the microwave popcorn market. the popcorn maker is a relatively new home appliance, and its popularity is increasing because it offers the opportunity to add flavors of the consumer own choice and to choose healthy-eating popcorn styles. popped sorghum is popular as a snack in india. the popped sorghum is similar to popcorn, but the puffs are smaller. recipes for popping sorghum by microwave, in a pot, etc., are readily available online. nutritional value air-popped popcorn (no salt or other additives) is 4% water, 78% carbohydrates (including 15% dietary fiber), 12% protein, and 4% fat (table). in a 100 gram reference amount, popcorn provides 382 calories and is a rich source (20% or more of the daily value, dv) of riboflavin (25% dv) and several dietary minerals, particularly manganese, phosphorus, and zinc (36-45% dv). b vitamins and other minerals are in appreciable amounts (table). saturated fat movie theaters commonly use coconut oil to pop the corn, and then top it with butter or margarine. movie theater popcorn contains large amounts of saturated fats and sodium due to its method of preparation. phytochemicals sorghum grains can be popped to form popcorn. all sorghums contain phenolic acids, and most contain flavonoids. sorghum grains are one of the highest food sources of the flavonoid proanthocyanidin. health risks popcorn is included on the list of foods that the american academy of pediatrics recommends not serving to children under four, because of the risk of choking. microwaveable popcorn represents a special case, since it is designed to be cooked along with its various flavoring agents. one of these formerly common artificial-butter flavorants, diacetyl, has been implicated in causing respiratory illnesses in microwave popcorn factory workers, also known as "popcorn lung". major manufacturers in the united states have stopped using this chemical, including orville redenbacher , act ii, pop secret and jolly time. other uses popcorn, threaded onto a string, is used as a wall or christmas tree decoration in some parts of north america, as well as on the balkan peninsula. some shipping companies have experimented with using popcorn as a biodegradable replacement for expanded polystyrene packing material. however, popcorn has numerous undesirable properties as a packing material, including attractiveness to pests, flammability, and a higher cost and greater density than expanded polystyrene. a more processed form of expanded corn foam has been developed to overcome some of these limitations, forming starch-based foam peanuts. see also references further reading hallauer, arnel r. (2001). specialty corns. crc press. isbn 978-0-8493-2377-5. lusas, edmund w.;

```
rooney, lloyd w. (2001). snack foods processing. crc press. isbn
978-1-56676-932-7. smith, andrew f. (1999). popped culture: the social history
of popcorn in america. university of south carolina press. isbn
978-1-57003-300-1.'
```

[28]:
```
# Break up single string into separate sentences
wiki_blob = TextBlob(wiki_text_clean)
len(wiki_blob.sentences)
```

[28]:
```
126
```

[29]:
```
# Only look at first 5 sentences
my_sentences = wiki_blob.sentences[0:5]
my_sentences
```

[29]:
```
[Sentence("popcorn (also called popped corn, popcorns, or pop-corn) is a variety
of corn kernel which expands and puffs up when heated."),
 Sentence("the term also refers to the snack food produced by the expansion."),
 Sentence("it is one of the oldest snacks, with evidence of popcorn dating back
thousands of years in the americas."),
 Sentence("it is commonly eaten salted, sweetened, or with artificial
flavorings."),
 Sentence("a popcorn kernel strong hull contains the seed hard, starchy shell
endosperm with 14-20% moisture, which turns to steam as the kernel is heated.")]
```

[30]:
```
# Convert text blob sentences to strings
my_sentences_str = [ str(x) for x in my_sentences ]
my_sentences_str
```

[30]:
```
['popcorn (also called popped corn, popcorns, or pop-corn) is a variety of corn
kernel which expands and puffs up when heated.',
 'the term also refers to the snack food produced by the expansion.',
 'it is one of the oldest snacks, with evidence of popcorn dating back thousands
of years in the americas.',
 'it is commonly eaten salted, sweetened, or with artificial flavorings.',
 'a popcorn kernel strong hull contains the seed hard, starchy shell endosperm
with 14-20% moisture, which turns to steam as the kernel is heated.']
```

[31]:
```
# Perform the TF-IDF Vectorization
tf_idf_vec = TfidfVectorizer(stop_words = 'english')
tf_idf_pop = tf_idf_vec.fit_transform(my_sentences_str)
tf_idf_pop.shape
```

[31]:
```
(5, 43)
```

[32]:
```
tf_idf_pop
```

[32]:
```
<Compressed Sparse Row sparse matrix of dtype 'float64'
        with 47 stored elements and shape (5, 43)>
```

[33]:
```
tf_idf_pop.transpose().shape
```

[33]:
```
(43, 5)
```

[34]:
```
tf_idf_vec.get_feature_names_out()
```

[34]:
```
array(['14', '20', 'americas', 'artificial', 'called', 'commonly',
       'contains', 'corn', 'dating', 'eaten', 'endosperm', 'evidence',
       'expands', 'expansion', 'flavorings', 'food', 'hard', 'heated',
       'hull', 'kernel', 'moisture', 'oldest', 'pop', 'popcorn',
       'popcorns', 'popped', 'produced', 'puffs', 'refers', 'salted',
       'seed', 'shell', 'snack', 'snacks', 'starchy', 'steam', 'strong',
       'sweetened', 'term', 'thousands', 'turns', 'variety', 'years'],
      dtype=object)
```

[35]:
```
# Print out results in a dataframe
tf_df = pd.DataFrame(tf_idf_pop.toarray(), columns = tf_idf_vec.
 ↪get_feature_names_out())
tf_df.transpose()
```

[35]:

	0	1	2	3	4
14	0.000000	0.000000	0.000000	0.000000	0.244682
20	0.000000	0.000000	0.000000	0.000000	0.244682
americas	0.000000	0.000000	0.366408	0.000000	0.000000
artificial	0.000000	0.000000	0.000000	0.408248	0.000000
called	0.237354	0.000000	0.000000	0.000000	0.000000
commonly	0.000000	0.000000	0.000000	0.408248	0.000000
contains	0.000000	0.000000	0.000000	0.000000	0.244682
corn	0.712062	0.000000	0.000000	0.000000	0.000000
dating	0.000000	0.000000	0.366408	0.000000	0.000000
eaten	0.000000	0.000000	0.000000	0.408248	0.000000
endosperm	0.000000	0.000000	0.000000	0.000000	0.244682
evidence	0.000000	0.000000	0.366408	0.000000	0.000000
expands	0.237354	0.000000	0.000000	0.000000	0.000000
expansion	0.000000	0.408248	0.000000	0.000000	0.000000
flavorings	0.000000	0.000000	0.000000	0.408248	0.000000
food	0.000000	0.408248	0.000000	0.000000	0.000000
hard	0.000000	0.000000	0.000000	0.000000	0.244682
heated	0.191496	0.000000	0.000000	0.000000	0.197408
hull	0.000000	0.000000	0.000000	0.000000	0.244682
kernel	0.191496	0.000000	0.000000	0.000000	0.394815
moisture	0.000000	0.000000	0.000000	0.000000	0.244682
oldest	0.000000	0.000000	0.366408	0.000000	0.000000
pop	0.237354	0.000000	0.000000	0.000000	0.000000

```
popcorn     0.158959  0.000000  0.245388  0.000000  0.163866
popcorns    0.237354  0.000000  0.000000  0.000000  0.000000
popped      0.237354  0.000000  0.000000  0.000000  0.000000
produced    0.000000  0.408248  0.000000  0.000000  0.000000
puffs       0.237354  0.000000  0.000000  0.000000  0.000000
refers      0.000000  0.408248  0.000000  0.000000  0.000000
salted      0.000000  0.000000  0.000000  0.408248  0.000000
seed        0.000000  0.000000  0.000000  0.000000  0.244682
shell       0.000000  0.000000  0.000000  0.000000  0.244682
snack       0.000000  0.408248  0.000000  0.000000  0.000000
snacks      0.000000  0.000000  0.366408  0.000000  0.000000
starchy     0.000000  0.000000  0.000000  0.000000  0.244682
steam       0.000000  0.000000  0.000000  0.000000  0.244682
strong      0.000000  0.000000  0.000000  0.000000  0.244682
sweetened   0.000000  0.000000  0.000000  0.408248  0.000000
term        0.000000  0.408248  0.000000  0.000000  0.000000
thousands   0.000000  0.000000  0.366408  0.000000  0.000000
turns       0.000000  0.000000  0.000000  0.000000  0.244682
variety     0.237354  0.000000  0.000000  0.000000  0.000000
years       0.000000  0.000000  0.366408  0.000000  0.000000
```

[36]:
```
# Get a data frame with the TF-IDF values sorted for document 0
df = pd.DataFrame(tf_idf_pop[0].T.todense(), index=tf_idf_vec.
 ↪get_feature_names_out(), columns=["TF-IDF"])
df = df.sort_values('TF-IDF', ascending=False)
df[:5]
```

[36]:
```
            TF-IDF
corn      0.712062
called    0.237354
expands   0.237354
pop       0.237354
popcorns  0.237354
```

[37]:
```
tf_df.T[[0]].sort_values([0], ascending=False)[:5]
```

[37]:
```
                0
corn      0.712062
called    0.237354
expands   0.237354
pop       0.237354
popcorns  0.237354
```

[38]:
```
(
tf_idf_pop[0]
.T
.todense()
```

```
)
```

[38]:
```
matrix([[0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.23735402],
        [0.        ],
        [0.        ],
        [0.71206206],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.23735402],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.19149573],
        [0.        ],
        [0.19149573],
        [0.        ],
        [0.        ],
        [0.23735402],
        [0.15895875],
        [0.23735402],
        [0.23735402],
        [0.        ],
        [0.23735402],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.        ],
        [0.23735402],
        [0.        ]])
```