3c-K.Nearest.Neighbors

November 8, 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 TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.neighbors import NearestNeighbors

pd.options.display.max_columns = 100
```

0.0.1 Gary Example

[2]:

```
my_df = pd.DataFrame()
my_df["names"] = ['Amantha', 'Brendon', 'Nate', 'Sam', 'Betty', 'Christine',␣
  ↪'Gin', 'Ken', 'Susy']
my_df["ages"] = [ 19, 23, 24, 30, 16, 18, 22, 18, 15 ]
my_df["genders_txt"] = "female male male male female female female male female".
  ↪split()
my_df["genders"] = [ 1, 0, 0, 0, 1, 1, 1, 0, 1 ]
my_df["music_band_txt"] = "Coldplay Coldplay LinkinPark LinkinPark Coldplay␣
  ↪LinkinPark LinkinPark Coldplay Coldplay".split()

my_df
```

[2]:

```
       names  ages genders_txt  genders music_band_txt
0    Amantha    19      female        1       Coldplay
1    Brendon    23        male        0       Coldplay
2       Nate    24        male        0     LinkinPark
3        Sam    30        male        0     LinkinPark
4      Betty    16      female        1       Coldplay
5  Christine    18      female        1     LinkinPark
6        Gin    22      female        1     LinkinPark
7        Ken    18        male        0       Coldplay
8       Susy    15      female        1       Coldplay
```

[3]:

```
my_df.select_dtypes("int")
```

```
[3]:    ages  genders
     0    19        1
     1    23        0
     2    24        0
     3    30        0
     4    16        1
     5    18        1
     6    22        1
     7    18        0
     8    15        1
```

Fit nearest neighbors

```
[4]: nn = NearestNeighbors().fit(my_df.select_dtypes("int"))
```

Get nearest neighbors distances

```
[5]: gary = pd.DataFrame( {"ages": [23], "genders": [0] } )
     gary
```

```
[5]:    ages  genders
     0    23        0
```

```
[6]: distances, indices = nn.kneighbors(
       X = gary,
       n_neighbors = 3,
     )
```

```
[7]: distances[0]**2
```

```
[7]: array([0., 1., 2.])
```

```
[8]: indices[0]
```

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

Get people matching index

```
[9]: my_df.iloc[indices[0]]
```

```
[9]:       names  ages genders_txt  genders music_band_txt
     1  Brendon    23        male        0       Coldplay
     2     Nate    24        male        0     LinkinPark
     6      Gin    22      female        1     LinkinPark
```

Vote

```
[10]: my_df.iloc[indices[0]]["music_band_txt"].mode()[0]
```

```
[10]: 'LinkinPark'
```

Repeat with K = all rows

```
[11]: d_i = nn.kneighbors(gary, n_neighbors = my_df.shape[0])
      distances, indices = np.array(d_i).reshape(2,9)
      distances**2, indices
```

```
[11]: (array([ 0.,  1.,  2., 17., 25., 26., 49., 50., 65.]),
       array([1., 2., 6., 0., 7., 5., 3., 4., 8.]))
```

```
[12]: (
        my_df
          .iloc[indices]
          .join( pd.DataFrame( { "distances^2": distances**2 }, index = indices ) )
      )
```

```
[12]:          names  ages genders_txt  genders music_band_txt  distances^2
      1.0    Brendon    23        male        0       Coldplay          0.0
      2.0       Nate    24        male        0     LinkinPark          1.0
      6.0        Gin    22      female        1     LinkinPark          2.0
      0.0    Amantha    19      female        1       Coldplay         17.0
      7.0        Ken    18        male        0       Coldplay         25.0
      5.0  Christine    18      female        1     LinkinPark         26.0
      3.0        Sam    30        male        0     LinkinPark         49.0
      4.0      Betty    16      female        1       Coldplay         50.0
      8.0       Susy    15      female        1       Coldplay         65.0
```

Display vote for various values of K ϵ { 1, 3, 5, 7, 9 }

```
[13]: for k in range(1,10,2):
        vote = my_df.iloc[indices]["music_band_txt"][:k].mode()[0]
        print(f"K = {k} : {vote}")
```

```
K = 1 : Coldplay
K = 3 : LinkinPark
K = 5 : Coldplay
K = 7 : LinkinPark
K = 9 : Coldplay
```

1 NLP

If our text data are unlabelled (as is often the case in NLP), we can use KNN to identify documents that are similar to a given document. In this example, our documents will be sentences and the given document will be the first sentence.

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

```
[15]: sentences_orig = [
        'Jen is a good student.',
        'Jen is also a great guitarist.',
        'Good students can sometimes be good guitarists',
      ]
      sentences_orig
```

```
[15]: ['Jen is a good student.',
       'Jen is also a great guitarist.',
       'Good students can sometimes be good guitarists']
```

2 Data Cleaning

We want to singularize guitarists and students.

```
[16]: sentence_last_tb = TextBlob(sentences_orig[-1]) # Make a textblob so that we␣
       ↪can singularize the word
      sentence_last_singular = [ x.singularize() for x in sentence_last_tb.words ] #␣
       ↪Singularize each word in the text
      sentence_last_clean = ' '.join(sentence_last_singular) # Join it together into␣
       ↪a single string
      sentence_last_clean
```

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

```
[17]: sentences_clean = sentences_orig[:2] + [sentence_last_clean]
      sentences_clean
```

```
[17]: ['Jen is a good student.',
       'Jen is also a great guitarist.',
       'Good student can sometime be good guitarist']
```

2.1 Bag of Words Using CountVectorizer

Perform the count transformation

```
[18]: vectorizer = CountVectorizer(stop_words='english')
      bow_matrix = vectorizer.fit_transform(sentences_clean)
```

```
[19]: type(bow_matrix), bow_matrix.shape
```

```
[19]: (scipy.sparse._csr.csr_matrix, (3, 5))
```

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

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

2.2 TF-IDF using BoW

Perform the TF-IDF transformation

[21]:
```
tf_idf_matrix = TfidfTransformer()
tf_idf_jen = tf_idf_matrix.fit_transform(bow_matrix)
```

[22]:
```
type(tf_idf_jen), tf_idf_jen.shape
```

```
[22]: (scipy.sparse._csr.csr_matrix, (3, 5))
```

[23]:
```
tf_idf_jen.toarray()
```

```
[23]: array([[0.57735027, 0.        , 0.        , 0.57735027, 0.57735027],
             [0.        , 0.68091856, 0.51785612, 0.51785612, 0.        ],
             [0.81649658, 0.        , 0.40824829, 0.        , 0.40824829]])
```

Print out results in a dataframe

[24]:
```
tf_df = pd.DataFrame(
  data = tf_idf_jen.toarray(),
  columns = vectorizer.get_feature_names_out(),
)
tf_df
```

```
[24]:        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
```

> Note: Converting a sparse matrix to a data frame is NOT something you will normally do, especially for large matrices.

2.3 K Nearest Neighbors

Fit nearest neighbors

[25]:
```
nn = NearestNeighbors().fit(tf_idf_jen)
```

Create the reference matrix from the tf_idf matrix

[26]:
```
sent0 = tf_idf_jen[0]
sent0.shape
```

```
[26]: (1, 5)
```

Or …

Create the reference matrix from the data frame

```
[27]: sent0 = np.array([tf_df.iloc[0]])
      sent0.shape
```

```
[27]: (1, 5)
```

Get nearest neighbors distances

```
[28]: distances, indices = nn.kneighbors(
        X = sent0,
        n_neighbors = 2,
      )
```

```
[29]: distances
```

```
[29]: array([[0.        , 0.76536686]])
```

```
[30]: indices
```

```
[30]: array([[0, 2]])
```

Pull out the original sentences given the indices.

```
[31]: # Using list comprehension
      [ x for i,x in enumerate(sentences_orig) if i in indices[0] ]
```

```
[31]: ['Jen is a good student.', 'Good students can sometimes be good guitarists']
```

```
[32]: # Converting to Numpy array
      np.array(sentences_orig)[indices]
```

```
[32]: array([['Jen is a good student.',
              'Good students can sometimes be good guitarists']], dtype='<U46')
```

3 Another Example - Using Wikipedia API

3.1 Get text and clean

Install Wikipedia API

```
[33]: %%capture
      !pip3 install wikipedia-api
```

```
[34]: import wikipediaapi
```

Pull out page from Wikipedia

```
[35]: topic = 'munchkin'
      wikip = wikipediaapi.Wikipedia('foobar')
      page_ex = wikip.page(topic)
      wiki_text = page_ex.text
```

```
wiki_text
```

[35]:

```
'A Munchkin is a native of the fictional Munchkin Country in the Oz books by
American author L. Frank Baum. They first appear in the classic children\'s
novel The Wonderful Wizard of Oz (1900) where they welcome Dorothy Gale to their
city in Oz. The Munchkins are described as being the same height as Dorothy and
they wear only shades of blue clothing, as blue is the Munchkins\' favorite
color. Blue is also the predominating color that officially represents the
eastern quadrant in the Land of Oz. The Munchkins have appeared in various
media, including the 1939 film The Wizard of Oz, as well as in various other
films and comedy acts.\n\nConcept\nWhile Baum may have written about it, there
are no surviving notes for the composition of The Wonderful Wizard of Oz. The
lack of this information has resulted in speculation of the term origins he used
in the book, which include the word Munchkin. Baum researcher Brian Attebery has
hypothesized that there might be a connection to the Münchner Kindl, the emblem
of the Bavarian city of Munich (spelled München in German). The symbol was
originally a 13th-century statue of a monk, looking down from the town hall in
Munich. Over the years, the image was reproduced many times, for instance as a
figure on beer steins, and eventually evolved into a child wearing a pointed
hood. Baum\'s family had German origins, suggesting that Baum could have seen
one such reproduction in his childhood. It is also possible that Munchkin came
from the German word Männchen, which means "mannikin" or "little figure". In
1900, Baum published a book about window displays in which he stressed the
importance of mannequins in attracting customers. Another possibility is a
connection to Baron Munchausen. This fictional character is based on a real
baron who told outrageous tall tales based on his military career. Like the
other Oz terms, the word Munchkin ends in a diminutive which in this case refers
to the size of the natives.\n\nLiterature\nOz Books by Frank Baum\nThe Munchkins
are first mentioned (quote shown) in an excerpt from chapter two of The
Wonderful Wizard of Oz, titled "The Council with the Munchkins". Dorothy
initially meets only three of them, along with the Good Witch of the North. The
rest of the Munchkins then come out of hiding and are shown to be grateful
towards Dorothy for killing their evil ruler the Wicked Witch of the East.
Dorothy later eventually finds the yellow brick road and along the way attends a
banquet held by a Munchkin man named Boq. Sometime in the book a background
story is also given about a "Munchkin maiden" (named Nimmie Amee in later
books), who was the former love interest of the Tin Woodman.\nBaum also included
the Munchkin characters in his later works as minor and major individual
characters. The Munchkin Jinjur is the main antagonist in Baum\'s second book
The Marvelous Land of Oz, where she seeks to overthrow the Scarecrow and take
over the Emerald City. Jinjur makes a brief appearance in the next book,
entitled Ozma of Oz, and is brought back in Baum\'s twelfth book, The Tin
Woodman of Oz. By this time, she is shown to be a more prominent character who
is helpful and friendly to Dorothy and her friends. Two other major Munchkin
characters also appear in The Tin Woodman of Oz: Tommy Kwikstep and Nimmie Amee.
The former appears in the story asking for a wish for running an errand for a
witch; the latter is the name given to the mystery "Munchkin maiden" from the
```

first book, who was the former lover of the Tin Woodman. More information is revealed that tells about the Tin Woodman\'s origin and their tragic love story. Lastly, the Munchkin Unc Nunkie appears in Baum\'s seventh book, The Patchwork Girl of Oz, where he is accidentally turned to stone. His Munchkin nephew Ojo successfully goes on a quest in search of an antidote while learning more about himself in the process.\n\nSubsequent Oz books\nL. Frank Baum died on 6 May 1919 after which other writers took up writing additional Oz stories. In some cases these books were written under Baum\'s name and included the Munchkins. There is at least one known Munchkin character that was created after Baum\'s death that appears as a major character. Zif is a Munchkin boy who appears in John R. Neill\'s first adaptation called The Royal Book of Oz. Zif is a student at the College of Art and Athletic Perfection; he is both respectful and resentful towards his teacher Wogglebog who considers Zif a "nobody or a nothing". The Munchkin characters that Baum had created in his lifetime also appear in these additional works.\n\nFilm and musicals\nEarly works (1902-1933)\nWhile the 1939 film is the most well known adaptation (see section below), it was not the first outside work to show the Munchkins in film or musical format. One of the first musical adaptations of Baum\'s books took place in 1902; it was also dubbed The Wizard of Oz. The Munchkins make their appearance in act one, called "The Storm", in which they are shown dancing around their maypole, not noticing that Dorothy\'s house has fallen to earth killing the Wicked Witch of the East. The first film adaptation of Baum\'s works, titled The Wonderful Wizard of Oz, was released in 1910, followed by three sequels. However, it was not until 1914 that Munchkin characters first appeared in film works. Ojo the Lucky and Unc Nunkie both appear in a film titled The Patchwork Girl of Oz (based on the book of the same name). This film stars American actress Violet MacMillan as Ojo and was produced by Baum.\n\n1939 film\nThe 1939 movie musical The Wizard of Oz was loosely based on Baum\'s novel. Notable differences of the Munchkins include their country name of Munchkinland and their clothes of many colors instead of an all-blue attire. In the musical, the Munchkins are portrayed by the thirty-odd members of the Singer Midgets, a European performing troupe made up of adult actors with dwarfism. Their numbers were swelled when a national talent search brought in a further ninety-four little men, women, and teenagers, with a few average-sized children were also included as background extras in order to make up the 124 characters requested by MGM.\nIn the musical, the Munchkins first appear when Dorothy and Toto arrive in the Land of Oz after her house lands on the Wicked Witch of the East. The Munchkins hide from all the commotion until Glinda the Good Witch arrives reassuring them that everything is okay. Dorothy tells them how she arrived in the Land of Oz (through a musical number) and the Munchkins celebrate. To make it official, a Barrister and a number of City Fathers insist to the Mayor of the Munchkin City that they must make sure that the Wicked Witch of the East is really dead before the celebration continues. The Coroner confirms this by saying that the witch is "not only merely dead" but is indeed "most sincerely dead" while showing a Certificate of Death. The Munchkins then celebrate further as Dorothy receives gifts from the "Lullaby League" and the "Lollipop Guild". Near the end of the song, the Wicked Witch of the West arrives, which causes the Munchkins to panic. After the Wicked Witch of the West leaves, Glinda tells Dorothy to follow the yellow brick road to the Emerald City as the Munchkins guide her out of Munchkinland.\nThe Munchkin actors have since not avoided controversy with alleged behavior behind the scenes. In a 1967 interview, Judy Garland referred to all of the Munchkins as "little drunks" who got intoxicated every night to the point where they had to be picked up in "butterfly nets". These accusations were denied as fabrications by fellow Munchkin Margaret Pellegrini, who said only "a couple of kids from Germany even drank beer". On 20 November 2007, the Munchkins were given a star on the Hollywood Walk of Fame. Seven of the surviving Munchkin actors from the film were present. As a result of the popularity of the 1939 film, the word "munchkin" has entered the English language as a reference to small children, persons with dwarfism, or anything of diminutive stature.\n\nActors and actresses\nThe following is a list of actors who portrayed the Munchkins in the 1939 film. Most of the dwarfs hired were acquired for MGM by Leo Singer, the proprietor of Singer\'s Midgets. A Daily Variety news story from 17 August 1938, stated 124 dwarves had been signed to play Munchkins; modern sources place the number either at 122 or 124. An additional dozen or so child actors were hired to make up for the shortage of dwarves. At least one Munchkin actor, Dale Paullin (stage name Paul Dale), did not make the final cut for the movie. Only two actors (Joseph Koziel and Frank Cucksey) used their actual voices for the dialogue exchanged with Dorothy where she is given the flowers. The rest of the voices, such as the "Munchkin chorus", were created by Pinto Colvig and Billy Bletcher with their voices recorded at a slow speed, which were subsequently sped-up when played back.\nIn 1989, author Stephen Cox researched, found, and wrote about the surviving Munchkin actors fifty years after they made the film. He wrote about them in his book, The Munchkins Remember (1989, E.P. Dutton), which was later revised as The Munchkins of Oz (Cumberland House), and his book remained in print for nearly two decades. When he wrote the book, 33 of the actors with dwarfism who appeared in the film were still alive and were interviewed. Several of them outlived all the major cast, as well as the original Tin Man Buddy Ebsen. Jerry Maren, who played the green "Lollipop Guild" member, was the last living adult Munchkin actor. Maren was the only Munchkin alive when the film\'s longest living cast member, Shep Houghton, an extra, died in 2016.\n\nNotes: Some of the information presented in the table below may never be complete as Social Security records remain sparse prior to the mid-twentieth century. Stage names and/or aliases are present in italics and quotation marks.\n\nChild actresses\nAbout a dozen children of average height were hired so they could be used for background fill. Sources differ on the number of children used for these roles ranging anywhere from 10 to 12. The names used for the women are maiden names with known aliases present in italics and quotation marks.\nAs of 2023, at least three "child munchkins" are known to be living.\n\nLater works (1940-1989)\nThe 1939 film was adapted into a musical that was released in 1942 that includes the Munchkin characters. The events that take place mirror the film including the song "Ding-Dong! The Witch Is Dead". Twenty-seven years later an animated film called The Wonderful Land of Oz was made featuring Jinjur as a main antagonist.\n\nOther works\nThe Munchkins appeared in The Wiz and were played by children and teenagers. (1978)\nThe

```
Munchkins appear at the end of Return to Oz. They are seen celebrating
Dorothy\'s return after defeating the Nome King and are present at Princess
Ozma\'s coronation. Tommy Kwikstep was also seen there. (1985)\nIn The Muppets\'
Wizard of Oz, the Munchkins were played by Rizzo the Rat (who portrayed the
"Mayor of Munchkinland") and his fellow rats. (2005)\nIn Strawberry Shortcake,
more specifically the 2003 cartoon, the fourth season contains an episode called
Berry Brick Road that involves a story where Strawberry Shortcake gets whisked
from her home. When she lands, she is greeted by three Munchkins that call
themselves the Berrykins (after a feylike being from the 1980s cartoon), were
tormented by the Wicked Witch of the West, thank Strawberry Shortcake for
knocking out the Wicked Witch of the West (which she only did by landing nearby)
and pressure her into stealing the latter\'s magic slippers (which she later
uses to return to her home) as a reward. She later returns to Oz to teach the
trio a lesson about caring for the environment. The Berrykins do not sing as
much as their people had in the original version, and they and the other
Munchkins look very small; however, the Berrykins specifically look just like
Blueberry Muffin, Rainbow Sherbet, and Lemon Meringue. (2007)\nThe Munchkins
appeared in Dorothy and the Witches of Oz. The Munchkins were first seen in the
battle against the Wicked Witch of the West\'s forces in Oz. They were later
brought to Earth by Glinda in order to combat the forces of the Wicked Witch of
the West. (2012)\nThe Munchkins appear in Oz the Great and Powerful. They
alongside the Quadlings and the Tinkers as inhabitants of Glinda\'s
protectorate. Although the film is not otherwise a musical, the Munchkins sing
and dance much as they do in the 1939 film. (2013)\nThe Munchkins appear in more
than one skit on Mad TV where the 1939 film is parodied. The actors are played
by people with dwarfism.\nThe Munchkins appear in the television series Once
Upon a Time. Not much is known about them, but they seem to be similar to the
Dwarves in the Enchanted forest as Zelena originally thought that Sneezy was a
Munchkin. Also, Regina Mills once mistakenly referred to the Seven Dwarfs as
Munchkins.\nThe Munchkins appear in Dorothy and the Wizard of Oz with the "Mayor
of Munchkinland" voiced by Bill Fagerbakke and the background Munchkins voiced
by Steven Blum and Jessica DiCicco. Ojo, Dr. Pipt, the Lollipop Guild, and the
Lullaby League are also featured. Also, Smith & Tinker are depicted as Munchkins
in this show.\n\nExplanatory notes\nReferences\nCitations\nWorks cited\nExternal
links\n The dictionary definition of munchkin at Wiktionary'
```

Replace newline chars with spaces before doing any processing. Strip the ' and "s" from possessives.

[36]:
```
wiki_text_clean = (
  wiki_text
  .replace("\n"," ")
  .replace("\'s",'')
  .replace('\'','')
  .replace("(", "")
  .replace(")", "")
  .replace('"', "")
)
```

```
wiki_text_clean
```

[36]:

```
'A Munchkin is a native of the fictional Munchkin Country in the Oz books by
American author L. Frank Baum. They first appear in the classic children novel
The Wonderful Wizard of Oz 1900 where they welcome Dorothy Gale to their city in
Oz. The Munchkins are described as being the same height as Dorothy and they
wear only shades of blue clothing, as blue is the Munchkins favorite color. Blue
is also the predominating color that officially represents the eastern quadrant
in the Land of Oz. The Munchkins have appeared in various media, including the
1939 film The Wizard of Oz, as well as in various other films and comedy acts.
Concept While Baum may have written about it, there are no surviving notes for
the composition of The Wonderful Wizard of Oz. The lack of this information has
resulted in speculation of the term origins he used in the book, which include
the word Munchkin. Baum researcher Brian Attebery has hypothesized that there
might be a connection to the Münchner Kindl, the emblem of the Bavarian city of
Munich spelled München in German. The symbol was originally a 13th-century
statue of a monk, looking down from the town hall in Munich. Over the years, the
image was reproduced many times, for instance as a figure on beer steins, and
eventually evolved into a child wearing a pointed hood. Baum family had German
origins, suggesting that Baum could have seen one such reproduction in his
childhood. It is also possible that Munchkin came from the German word Männchen,
which means mannikin or little figure. In 1900, Baum published a book about
window displays in which he stressed the importance of mannequins in attracting
customers. Another possibility is a connection to Baron Munchausen. This
fictional character is based on a real baron who told outrageous tall tales
based on his military career. Like the other Oz terms, the word Munchkin ends in
a diminutive which in this case refers to the size of the natives.  Literature
Oz Books by Frank Baum The Munchkins are first mentioned quote shown in an
excerpt from chapter two of The Wonderful Wizard of Oz, titled The Council with
the Munchkins. Dorothy initially meets only three of them, along with the Good
Witch of the North. The rest of the Munchkins then come out of hiding and are
shown to be grateful towards Dorothy for killing their evil ruler the Wicked
Witch of the East. Dorothy later eventually finds the yellow brick road and
along the way attends a banquet held by a Munchkin man named Boq. Sometime in
the book a background story is also given about a Munchkin maiden named Nimmie
Amee in later books, who was the former love interest of the Tin Woodman. Baum
also included the Munchkin characters in his later works as minor and major
individual characters. The Munchkin Jinjur is the main antagonist in Baum second
book The Marvelous Land of Oz, where she seeks to overthrow the Scarecrow and
take over the Emerald City. Jinjur makes a brief appearance in the next book,
entitled Ozma of Oz, and is brought back in Baum twelfth book, The Tin Woodman
of Oz. By this time, she is shown to be a more prominent character who is
helpful and friendly to Dorothy and her friends. Two other major Munchkin
characters also appear in The Tin Woodman of Oz: Tommy Kwikstep and Nimmie Amee.
The former appears in the story asking for a wish for running an errand for a
witch; the latter is the name given to the mystery Munchkin maiden from the
first book, who was the former lover of the Tin Woodman. More information is
```

revealed that tells about the Tin Woodman origin and their tragic love story. Lastly, the Munchkin Unc Nunkie appears in Baum seventh book, The Patchwork Girl of Oz, where he is accidentally turned to stone. His Munchkin nephew Ojo successfully goes on a quest in search of an antidote while learning more about himself in the process. Subsequent Oz books L. Frank Baum died on 6 May 1919 after which other writers took up writing additional Oz stories. In some cases these books were written under Baum name and included the Munchkins. There is at least one known Munchkin character that was created after Baum death that appears as a major character. Zif is a Munchkin boy who appears in John R. Neill first adaptation called The Royal Book of Oz. Zif is a student at the College of Art and Athletic Perfection; he is both respectful and resentful towards his teacher Wogglebog who considers Zif a nobody or a nothing. The Munchkin characters that Baum had created in his lifetime also appear in these additional works. Film and musicals Early works 1902-1933 While the 1939 film is the most well known adaptation see section below, it was not the first outside work to show the Munchkins in film or musical format. One of the first musical adaptations of Baum books took place in 1902; it was also dubbed The Wizard of Oz. The Munchkins make their appearance in act one, called The Storm, in which they are shown dancing around their maypole, not noticing that Dorothy house has fallen to earth killing the Wicked Witch of the East. The first film adaptation of Baum works, titled The Wonderful Wizard of Oz, was released in 1910, followed by three sequels. However, it was not until 1914 that Munchkin characters first appeared in film works. Ojo the Lucky and Unc Nunkie both appear in a film titled The Patchwork Girl of Oz based on the book of the same name. This film stars American actress Violet MacMillan as Ojo and was produced by Baum. 1939 film The 1939 movie musical The Wizard of Oz was loosely based on Baum novel. Notable differences of the Munchkins include their country name of Munchkinland and their clothes of many colors instead of an all-blue attire. In the musical, the Munchkins are portrayed by the thirty-odd members of the Singer Midgets, a European performing troupe made up of adult actors with dwarfism. Their numbers were swelled when a national talent search brought in a further ninety-four little men, women, and teenagers, with a few average-sized children were also included as background extras in order to make up the 124 characters requested by MGM. In the musical, the Munchkins first appear when Dorothy and Toto arrive in the Land of Oz after her house lands on the Wicked Witch of the East. The Munchkins hide from all the commotion until Glinda the Good Witch arrives reassuring them that everything is okay. Dorothy tells them how she arrived in the Land of Oz through a musical number and the Munchkins celebrate. To make it official, a Barrister and a number of City Fathers insist to the Mayor of the Munchkin City that they must make sure that the Wicked Witch of the East is really dead before the celebration continues. The Coroner confirms this by saying that the witch is not only merely dead but is indeed most sincerely dead while showing a Certificate of Death. The Munchkins then celebrate further as Dorothy receives gifts from the Lullaby League and the Lollipop Guild. Near the end of the song, the Wicked Witch of the West arrives, which causes the Munchkins to panic. After the Wicked Witch of the West leaves, Glinda tells Dorothy to follow the yellow brick road to the Emerald City as the Munchkins guide her out of Munchkinland. The Munchkin actors have since not avoided controversy with alleged behavior behind the scenes. In a 1967 interview, Judy Garland referred to all of the Munchkins as little drunks who got intoxicated every night to the point where they had to be picked up in butterfly nets. These accusations were denied as fabrications by fellow Munchkin Margaret Pellegrini, who said only a couple of kids from Germany even drank beer. On 20 November 2007, the Munchkins were given a star on the Hollywood Walk of Fame. Seven of the surviving Munchkin actors from the film were present. As a result of the popularity of the 1939 film, the word munchkin has entered the English language as a reference to small children, persons with dwarfism, or anything of diminutive stature. Actors and actresses The following is a list of actors who portrayed the Munchkins in the 1939 film. Most of the dwarfs hired were acquired for MGM by Leo Singer, the proprietor of Singer Midgets. A Daily Variety news story from 17 August 1938, stated 124 dwarves had been signed to play Munchkins; modern sources place the number either at 122 or 124. An additional dozen or so child actors were hired to make up for the shortage of dwarves. At least one Munchkin actor, Dale Paullin stage name Paul Dale, did not make the final cut for the movie. Only two actors Joseph Koziel and Frank Cucksey used their actual voices for the dialogue exchanged with Dorothy where she is given the flowers. The rest of the voices, such as the Munchkin chorus, were created by Pinto Colvig and Billy Bletcher with their voices recorded at a slow speed, which were subsequently sped-up when played back. In 1989, author Stephen Cox researched, found, and wrote about the surviving Munchkin actors fifty years after they made the film. He wrote about them in his book, The Munchkins Remember 1989, E.P. Dutton, which was later revised as The Munchkins of Oz Cumberland House, and his book remained in print for nearly two decades. When he wrote the book, 33 of the actors with dwarfism who appeared in the film were still alive and were interviewed. Several of them outlived all the major cast, as well as the original Tin Man Buddy Ebsen. Jerry Maren, who played the green Lollipop Guild member, was the last living adult Munchkin actor. Maren was the only Munchkin alive when the film longest living cast member, Shep Houghton, an extra, died in 2016. Notes: Some of the information presented in the table below may never be complete as Social Security records remain sparse prior to the mid-twentieth century. Stage names and/or aliases are present in italics and quotation marks. Child actresses About a dozen children of average height were hired so they could be used for background fill. Sources differ on the number of children used for these roles ranging anywhere from 10 to 12. The names used for the women are maiden names with known aliases present in italics and quotation marks. As of 2023, at least three child munchkins are known to be living. Later works 1940-1989 The 1939 film was adapted into a musical that was released in 1942 that includes the Munchkin characters. The events that take place mirror the film including the song Ding-Dong! The Witch Is Dead. Twenty-seven years later an animated film called The Wonderful Land of Oz was made featuring Jinjur as a main antagonist. Other works The Munchkins appeared in The Wiz and were played by children and teenagers. 1978 The Munchkins appear at the end of Return to Oz. They are seen celebrating Dorothy return after defeating the Nome King and are present at Princess Ozma coronation. Tommy Kwikstep was also seen there. 1985 In

```
The Muppets Wizard of Oz, the Munchkins were played by Rizzo the Rat who
portrayed the Mayor of Munchkinland and his fellow rats. 2005 In Strawberry
Shortcake, more specifically the 2003 cartoon, the fourth season contains an
episode called Berry Brick Road that involves a story where Strawberry Shortcake
gets whisked from her home. When she lands, she is greeted by three Munchkins
that call themselves the Berrykins after a feylike being from the 1980s cartoon,
were tormented by the Wicked Witch of the West, thank Strawberry Shortcake for
knocking out the Wicked Witch of the West which she only did by landing nearby
and pressure her into stealing the latter magic slippers which she later uses to
return to her home as a reward. She later returns to Oz to teach the trio a
lesson about caring for the environment. The Berrykins do not sing as much as
their people had in the original version, and they and the other Munchkins look
very small; however, the Berrykins specifically look just like Blueberry Muffin,
Rainbow Sherbet, and Lemon Meringue. 2007 The Munchkins appeared in Dorothy and
the Witches of Oz. The Munchkins were first seen in the battle against the
Wicked Witch of the West forces in Oz. They were later brought to Earth by
Glinda in order to combat the forces of the Wicked Witch of the West. 2012 The
Munchkins appear in Oz the Great and Powerful. They alongside the Quadlings and
the Tinkers as inhabitants of Glinda protectorate. Although the film is not
otherwise a musical, the Munchkins sing and dance much as they do in the 1939
film. 2013 The Munchkins appear in more than one skit on Mad TV where the 1939
film is parodied. The actors are played by people with dwarfism. The Munchkins
appear in the television series Once Upon a Time. Not much is known about them,
but they seem to be similar to the Dwarves in the Enchanted forest as Zelena
originally thought that Sneezy was a Munchkin. Also, Regina Mills once
mistakenly referred to the Seven Dwarfs as Munchkins. The Munchkins appear in
Dorothy and the Wizard of Oz with the Mayor of Munchkinland voiced by Bill
Fagerbakke and the background Munchkins voiced by Steven Blum and Jessica
DiCicco. Ojo, Dr. Pipt, the Lollipop Guild, and the Lullaby League are also
featured. Also, Smith & Tinker are depicted as Munchkins in this show.
Explanatory notes References Citations Works cited External links  The
dictionary definition of munchkin at Wiktionary'
```

Convert to textblob

```
[37]: wiki_blob = TextBlob(wiki_text_clean)
```

Only look at first 5 sentences

```
[49]: my_sentences = wiki_blob.sentences[0:20]
      my_sentences
```

```
[49]: [Sentence("A Munchkin is a native of the fictional Munchkin Country in the Oz
books by American author L. Frank Baum."),
 Sentence("They first appear in the classic children novel The Wonderful Wizard
of Oz 1900 where they welcome Dorothy Gale to their city in Oz."),
 Sentence("The Munchkins are described as being the same height as Dorothy and
they wear only shades of blue clothing, as blue is the Munchkins favorite
color."),
 Sentence("Blue is also the predominating color that officially represents the
eastern quadrant in the Land of Oz."),
 Sentence("The Munchkins have appeared in various media, including the 1939 film
The Wizard of Oz, as well as in various other films and comedy acts."),
 Sentence("Concept While Baum may have written about it, there are no surviving
notes for the composition of The Wonderful Wizard of Oz."),
 Sentence("The lack of this information has resulted in speculation of the term
origins he used in the book, which include the word Munchkin."),
 Sentence("Baum researcher Brian Attebery has hypothesized that there might be a
connection to the Münchner Kindl, the emblem of the Bavarian city of Munich
spelled München in German."),
 Sentence("The symbol was originally a 13th-century statue of a monk, looking
down from the town hall in Munich."),
 Sentence("Over the years, the image was reproduced many times, for instance as
a figure on beer steins, and eventually evolved into a child wearing a pointed
hood."),
 Sentence("Baum family had German origins, suggesting that Baum could have seen
one such reproduction in his childhood."),
 Sentence("It is also possible that Munchkin came from the German word Männchen,
which means mannikin or little figure."),
 Sentence("In 1900, Baum published a book about window displays in which he
stressed the importance of mannequins in attracting customers."),
 Sentence("Another possibility is a connection to Baron Munchausen."),
 Sentence("This fictional character is based on a real baron who told outrageous
tall tales based on his military career."),
 Sentence("Like the other Oz terms, the word Munchkin ends in a diminutive which
in this case refers to the size of the natives."),
 Sentence("Literature Oz Books by Frank Baum The Munchkins are first mentioned
quote shown in an excerpt from chapter two of The Wonderful Wizard of Oz, titled
The Council with the Munchkins."),
 Sentence("Dorothy initially meets only three of them, along with the Good Witch
of the North."),
 Sentence("The rest of the Munchkins then come out of hiding and are shown to be
grateful towards Dorothy for killing their evil ruler the Wicked Witch of the
East."),
 Sentence("Dorothy later eventually finds the yellow brick road and along the
way attends a banquet held by a Munchkin man named Boq.")]
```

[50]:

```python
len(wiki_blob.sentences)
```

[50]:

```
109
```

Singularize and convert back to string

[51]:

```python
for i, sentence in enumerate(my_sentences):
  sing = [x.singularize() for x in sentence.words]
  my_sentences[i] = ' '.join(sing)
```

```
my_sentences
```

```
[51]: ['A Munchkin is a native of the fictional Munchkin Country in the Oz book by
      American author L Frank Baum',
       'They first appear in the classic child novel The Wonderful Wizard of Oz 1900
      where they welcome Dorothy Gale to their city in Oz',
       'The Munchkin are described a being the same height a Dorothy and they wear
      only shade of blue clothing a blue is the Munchkin favorite color',
       'Blue is also the predominating color that officially represent the eastern
      quadrant in the Land of Oz',
       'The Munchkin have appeared in variou medium including the 1939 film The Wizard
      of Oz a well a in variou other film and comedy act',
       'Concept While Baum may have written about it there are no surviving note for
      the composition of The Wonderful Wizard of Oz',
       'The lack of thi information ha resulted in speculation of the term origin he
      used in the book which include the word Munchkin',
       'Baum researcher Brian Attebery ha hypothesized that there might be a
      connection to the Münchner Kindl the emblem of the Bavarian city of Munich
      spelled München in German',
       'The symbol wa originally a 13th-century statue of a monk looking down from the
      town hall in Munich',
       'Over the year the image wa reproduced many time for instance a a figure on
      beer stein and eventually evolved into a child wearing a pointed hood',
       'Baum family had German origin suggesting that Baum could have seen one such
      reproduction in hi childhood',
       'It is also possible that Munchkin came from the German word Männchen which
      mean mannikin or little figure',
       'In 1900 Baum published a book about window display in which he stressed the
      importance of mannequin in attracting customer',
       'Another possibility is a connection to Baron Munchausen',
       'Thi fictional character is based on a real baron who told outrageou tall tale
      based on hi military career',
       'Like the other Oz term the word Munchkin end in a diminutive which in thi case
      refer to the size of the native',
       'Literature Oz Book by Frank Baum The Munchkin are first mentioned quote shown
      in an excerpt from chapter two of The Wonderful Wizard of Oz titled The Council
      with the Munchkin',
       'Dorothy initially meet only three of them along with the Good Witch of the
      North',
       'The rest of the Munchkin then come out of hiding and are shown to be grateful
      toward Dorothy for killing their evil ruler the Wicked Witch of the East',
       'Dorothy later eventually find the yellow brick road and along the way attend a
      banquet held by a Munchkin man named Boq']
```

3.2 TF-IDF without using BoW

Perform the TF-IDF Vectorization

[52]:

```
tf_idf_matrix = TfidfVectorizer(stop_words = 'english')
tf_idf = tf_idf_matrix.fit_transform(my_sentences)
```

[53]:

```
tf_idf.shape
```

[53]:

```
(20, 166)
```

Print out results in a data frame

[54]:

```
results_df = pd.DataFrame(
  data = tf_idf.toarray(),
  columns = tf_idf_matrix.get_feature_names_out()
)
results_df.transpose()
```

[54]:

```
                 0         1    2    3         4         5         6    7   \
13th      0.000000  0.000000  0.0  0.0  0.000000  0.000000  0.000000  0.0
1900      0.000000  0.273863  0.0  0.0  0.000000  0.000000  0.000000  0.0
1939      0.000000  0.000000  0.0  0.0  0.256986  0.000000  0.000000  0.0
act       0.000000  0.000000  0.0  0.0  0.256986  0.000000  0.000000  0.0
american  0.361256  0.000000  0.0  0.0  0.000000  0.000000  0.000000  0.0
…             …       …  …  …       …       …       …  …
wonderful 0.000000  0.247119  0.0  0.0  0.000000  0.302110  0.000000  0.0
word      0.000000  0.000000  0.0  0.0  0.000000  0.000000  0.246257  0.0
written   0.000000  0.000000  0.0  0.0  0.000000  0.380887  0.000000  0.0
year      0.000000  0.000000  0.0  0.0  0.000000  0.000000  0.000000  0.0
yellow    0.000000  0.000000  0.0  0.0  0.000000  0.000000  0.000000  0.0

                 8       9    10        11        12   13   14        15  \
13th      0.307943  0.0000  0.0  0.000000  0.000000  0.0  0.0  0.000000
1900      0.000000  0.0000  0.0  0.000000  0.282341  0.0  0.0  0.000000
1939      0.000000  0.0000  0.0  0.000000  0.000000  0.0  0.0  0.000000
act       0.000000  0.0000  0.0  0.000000  0.000000  0.0  0.0  0.000000
american  0.000000  0.0000  0.0  0.000000  0.000000  0.0  0.0  0.000000
…             …     …  …       …       …  …  …       …
wonderful 0.000000  0.0000  0.0  0.000000  0.000000  0.0  0.0  0.000000
word      0.000000  0.0000  0.0  0.275297  0.000000  0.0  0.0  0.258464
written   0.000000  0.0000  0.0  0.000000  0.000000  0.0  0.0  0.000000
year      0.000000  0.2664  0.0  0.000000  0.000000  0.0  0.0  0.000000
yellow    0.000000  0.0000  0.0  0.000000  0.000000  0.0  0.0  0.000000

                 16   17   18        19
13th      0.000000  0.0  0.0  0.000000
1900      0.000000  0.0  0.0  0.000000
1939      0.000000  0.0  0.0  0.000000
act       0.000000  0.0  0.0  0.000000
american  0.000000  0.0  0.0  0.000000
…             …  …  …       …
```

```
wonderful  0.219328  0.0  0.0  0.000000
word       0.000000  0.0  0.0  0.000000
written    0.000000  0.0  0.0  0.000000
year       0.000000  0.0  0.0  0.000000
yellow     0.000000  0.0  0.0  0.282903

[166 rows x 20 columns]
```

3.3 K Nearest Neighbors

Fit nearest neighbors

```
[55]: nn = NearestNeighbors().fit(tf_idf)
```

Get nearest neighbors distances to first sentence

```
[57]: distances, indices = nn.kneighbors(
        X = tf_idf[0],
        n_neighbors = 3,
      )
```

```
[58]: distances
```

```
[58]: array([[0.        , 1.14391441, 1.26882977]])
```

```
[59]: indices
```

```
[59]: array([[ 0, 16, 15]])
```

```
[60]: np.array(my_sentences)[indices]
```

```
[60]: array([['A Munchkin is a native of the fictional Munchkin Country in the Oz book
      by American author L Frank Baum',
              'Literature Oz Book by Frank Baum The Munchkin are first mentioned quote
      shown in an excerpt from chapter two of The Wonderful Wizard of Oz titled The
      Council with the Munchkin',
              'Like the other Oz term the word Munchkin end in a diminutive which in
      thi case refer to the size of the native']],
            dtype='<U175')
```

```
[ ]:
```