data_transformation_2_text

December 30, 2018

In [1]:
```
import pandas as pd
categotical_feature = pd.Series(['sunny','cloudy', 'snowy', 'rainy', 'foggy'])
mapping = pd.get_dummies(categotical_feature)
mapping
```

Out[1]:

	cloudy	foggy	rainy	snowy	sunny
0	0	0	0	0	1
1	1	0	0	0	0
2	0	0	0	1	0
3	0	0	1	0	0
4	0	1	0	0	0

In [11]:
```
### Same thing but with sklearn
from sklearn.preprocessing import OneHotEncoder, LabelEncoder
le = LabelEncoder()
ohe = OneHotEncoder()
levels = ['sunny', 'cloudy', 'snowy', 'rainy', 'foggy']
fit_levels = le.fit_transform(levels)
levels_transformed = [ [level] for level in fit_levels]
ohe.fit(levels_transformed)
print(levels_transformed)
print(ohe.transform([le.transform(['sunny'])]).toarray())
print(ohe.transform([le.transform(['cloudy'])]).toarray())
```

```
[[4], [0], [3], [2], [1]]
[[0. 0. 0. 0. 1.]]
[[1. 0. 0. 0. 0.]]
```

```
d:\python\lib\site-packages\sklearn\preprocessing\_encoders.py:368: FutureWarning: The handling
If you want the future behaviour and silence this warning, you can specify "categories='auto'"
In case you used a LabelEncoder before this OneHotEncoder to convert the categories to integers
  warnings.warn(msg, FutureWarning)
```

In [12]:
```
### Text data
from sklearn.datasets import fetch_20newsgroups
categories = ['sci.med', 'sci.space']
twenty_sci_news = fetch_20newsgroups(categories=categories)
```

```
Downloading 20news dataset. This may take a few minutes.
Downloading dataset from https://ndownloader.figshare.com/files/5975967 (14 MB)
```

```
In [13]: print(twenty_sci_news.data[0])
```

```
From: flb@flb.optiplan.fi ("F.Baube[tm]")
Subject: Vandalizing the sky
X-Added: Forwarded by Space Digest
Organization: [via International Space University]
Original-Sender: isu@VACATION.VENARI.CS.CMU.EDU
Distribution: sci
Lines: 12

From: "Phil G. Fraering" <pgf@srl03.cacs.usl.edu>
>
> Finally: this isn't the Bronze Age, [..]
> please try to remember that there are more human activities than
> those practiced by the Warrior Caste, the Farming Caste, and the
> Priesthood.

Right, the Profiting Caste is blessed by God, and may
 freely blare its presence in the evening twilight ..

--
* Fred Baube (tm)
```

```
In [15]: twenty_sci_news.filenames
```

```
Out[15]: array(['C:\\Users\\Tomasz\\scikit_learn_data\\20news_home\\20news-bydate-train\\sci.sp
                'C:\\Users\\Tomasz\\scikit_learn_data\\20news_home\\20news-bydate-train\\sci.me
                'C:\\Users\\Tomasz\\scikit_learn_data\\20news_home\\20news-bydate-train\\sci.me
                ...,
                'C:\\Users\\Tomasz\\scikit_learn_data\\20news_home\\20news-bydate-train\\sci.sp
                'C:\\Users\\Tomasz\\scikit_learn_data\\20news_home\\20news-bydate-train\\sci.sp
                'C:\\Users\\Tomasz\\scikit_learn_data\\20news_home\\20news-bydate-train\\sci.me
               dtype='<U96')
```

```
In [16]: ### count words
         from sklearn.feature_extraction.text import CountVectorizer
         count_vect = CountVectorizer()
         word_count = count_vect.fit_transform(twenty_sci_news.data)
         word_count.shape
```

```
Out[16]: (1187, 25638)
```

```
In [18]: print(word_count[0])
```

```
(0, 10778)        1
(0, 23849)        1
(0, 9796)        1
(0, 12716)        1
(0, 18586)        1
(0, 13384)        1
(0, 5134)        1
(0, 10785)        1
(0, 15246)        1
(0, 11330)        1
(0, 5148)        1
(0, 13318)        1
(0, 18744)        1
(0, 20110)        1
(0, 18642)        1
(0, 3808)        2
(0, 10188)        1
(0, 6017)        3
(0, 24930)        1
(0, 18474)        1
(0, 23241)        1
(0, 23129)        1
(0, 3191)        1
(0, 12362)        1
(0, 15968)        1
:        :
(0, 7646)        1
(0, 24547)        1
(0, 24415)        1
(0, 13359)        1
(0, 20909)        1
(0, 17235)        1
(0, 24151)        1
(0, 13158)        1
(0, 24626)        1
(0, 17217)        1
(0, 8438)        1
(0, 21686)        2
(0, 5650)        3
(0, 10713)        1
(0, 3233)        1
(0, 21382)        1
(0, 23137)        7
(0, 24461)        1
(0, 22345)        1
(0, 23381)        2
(0, 4762)        2
(0, 10341)        1
```

```
  (0, 17170)        1
  (0, 10501)        2
  (0, 10827)        2
```

```
In [20]: word_list = count_vect.get_feature_names()
         for n in word_count[0].indices:
             print("Word: %s. Count: %s" % (word_list[n], word_count[0, n]))
```

```
Word: fred Count: 1
Word: twilight Count: 1
Word: evening Count: 1
Word: in Count: 1
Word: presence Count: 1
Word: its Count: 1
Word: blare Count: 1
Word: freely Count: 1
Word: may Count: 1
Word: god Count: 1
Word: blessed Count: 1
Word: is Count: 1
Word: profiting Count: 1
Word: right Count: 1
Word: priesthood Count: 1
Word: and Count: 2
Word: farming Count: 1
Word: caste Count: 3
Word: warrior Count: 1
Word: practiced Count: 1
Word: those Count: 1
Word: than Count: 1
Word: activities Count: 1
Word: human Count: 1
Word: more Count: 1
Word: are Count: 1
Word: there Count: 1
Word: that Count: 1
Word: remember Count: 1
Word: to Count: 1
Word: try Count: 1
Word: please Count: 1
Word: age Count: 1
Word: bronze Count: 1
Word: isn Count: 1
Word: this Count: 1
Word: finally Count: 1
Word: usl Count: 1
Word: cacs Count: 1
```

```
Word: srl03 Count: 1
Word: pgf Count: 1
Word: fraering Count: 1
Word: phil Count: 1
Word: 12 Count: 1
Word: lines Count: 1
Word: sci Count: 1
Word: distribution Count: 1
Word: edu Count: 2
Word: cmu Count: 1
Word: cs Count: 1
Word: venari Count: 1
Word: vacation Count: 1
Word: isu Count: 1
Word: sender Count: 1
Word: original Count: 1
Word: university Count: 1
Word: international Count: 1
Word: via Count: 1
Word: organization Count: 1
Word: digest Count: 1
Word: space Count: 2
Word: by Count: 3
Word: forwarded Count: 1
Word: added Count: 1
Word: sky Count: 1
Word: the Count: 7
Word: vandalizing Count: 1
Word: subject Count: 1
Word: tm Count: 2
Word: baube Count: 2
Word: fi Count: 1
Word: optiplan Count: 1
Word: flb Count: 2
Word: from Count: 2
```

```
In [25]: ### Word Frequency
         from sklearn.feature_extraction.text import TfidfVectorizer
         tf_vect = TfidfVectorizer(use_idf=False, norm='l1')
         word_freq = tf_vect.fit_transform(twenty_sci_news.data)
         word_list = tf_vect.get_feature_names()
         for n in word_freq[0].indices:
             print("Word: %s. Frequency: %0.3f" % (word_list[n], word_freq[0,n]))
```

```
Word: fred. Frequency: 0.011
Word: twilight. Frequency: 0.011
Word: evening. Frequency: 0.011
```

Word: in. Frequency: 0.011
Word: presence. Frequency: 0.011
Word: its. Frequency: 0.011
Word: blare. Frequency: 0.011
Word: freely. Frequency: 0.011
Word: may. Frequency: 0.011
Word: god. Frequency: 0.011
Word: blessed. Frequency: 0.011
Word: is. Frequency: 0.011
Word: profiting. Frequency: 0.011
Word: right. Frequency: 0.011
Word: priesthood. Frequency: 0.011
Word: and. Frequency: 0.022
Word: farming. Frequency: 0.011
Word: caste. Frequency: 0.033
Word: warrior. Frequency: 0.011
Word: practiced. Frequency: 0.011
Word: those. Frequency: 0.011
Word: than. Frequency: 0.011
Word: activities. Frequency: 0.011
Word: human. Frequency: 0.011
Word: more. Frequency: 0.011
Word: are. Frequency: 0.011
Word: there. Frequency: 0.011
Word: that. Frequency: 0.011
Word: remember. Frequency: 0.011
Word: to. Frequency: 0.011
Word: try. Frequency: 0.011
Word: please. Frequency: 0.011
Word: age. Frequency: 0.011
Word: bronze. Frequency: 0.011
Word: isn. Frequency: 0.011
Word: this. Frequency: 0.011
Word: finally. Frequency: 0.011
Word: usl. Frequency: 0.011
Word: cacs. Frequency: 0.011
Word: srl03. Frequency: 0.011
Word: pgf. Frequency: 0.011
Word: fraering. Frequency: 0.011
Word: phil. Frequency: 0.011
Word: 12. Frequency: 0.011
Word: lines. Frequency: 0.011
Word: sci. Frequency: 0.011
Word: distribution. Frequency: 0.011
Word: edu. Frequency: 0.022
Word: cmu. Frequency: 0.011
Word: cs. Frequency: 0.011
Word: venari. Frequency: 0.011

```
Word: vacation. Frequency: 0.011
Word: isu. Frequency: 0.011
Word: sender. Frequency: 0.011
Word: original. Frequency: 0.011
Word: university. Frequency: 0.011
Word: international. Frequency: 0.011
Word: via. Frequency: 0.011
Word: organization. Frequency: 0.011
Word: digest. Frequency: 0.011
Word: space. Frequency: 0.022
Word: by. Frequency: 0.033
Word: forwarded. Frequency: 0.011
Word: added. Frequency: 0.011
Word: sky. Frequency: 0.011
Word: the. Frequency: 0.077
Word: vandalizing. Frequency: 0.011
Word: subject. Frequency: 0.011
Word: tm. Frequency: 0.022
Word: baube. Frequency: 0.022
Word: fi. Frequency: 0.011
Word: optiplan. Frequency: 0.011
Word: flb. Frequency: 0.022
Word: from. Frequency: 0.022
```

In [28]:
```
## TFIDF algoritm short for term frequencyinverse document frequency, is a numerical
## word is to a document in a collection or corpus.[1] It is often used as a weightin
## retrieval, text mining, and user modeling. The tfidf value increases proportionall
## in the document and is offset by the number of documents in the corpus that contai
##fact that some words appear more frequently in general. Tfidf is one of the most po

from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_vect = TfidfVectorizer()
word_tfidf = tfidf_vect.fit_transform(twenty_sci_news.data)
word_list = tfidf_vect.get_feature_names()
for n in word_tfidf[0].indices:
    print("Word: %s. Value tf-idf: %0.3f" % (word_list[n], word_tfidf[0,n]))
```

```
Word: from. Value tf-idf: 0.043
Word: flb. Value tf-idf: 0.264
Word: optiplan. Value tf-idf: 0.132
Word: fi. Value tf-idf: 0.110
Word: baube. Value tf-idf: 0.264
Word: tm. Value tf-idf: 0.219
Word: subject. Value tf-idf: 0.022
Word: vandalizing. Value tf-idf: 0.103
Word: the. Value tf-idf: 0.158
Word: sky. Value tf-idf: 0.091
```

Word: added. Value tf-idf: 0.088
Word: forwarded. Value tf-idf: 0.096
Word: by. Value tf-idf: 0.120
Word: space. Value tf-idf: 0.098
Word: digest. Value tf-idf: 0.095
Word: organization. Value tf-idf: 0.022
Word: via. Value tf-idf: 0.083
Word: international. Value tf-idf: 0.081
Word: university. Value tf-idf: 0.045
Word: original. Value tf-idf: 0.085
Word: sender. Value tf-idf: 0.093
Word: isu. Value tf-idf: 0.099
Word: vacation. Value tf-idf: 0.099
Word: venari. Value tf-idf: 0.103
Word: cs. Value tf-idf: 0.055
Word: cmu. Value tf-idf: 0.081
Word: edu. Value tf-idf: 0.059
Word: distribution. Value tf-idf: 0.053
Word: sci. Value tf-idf: 0.067
Word: lines. Value tf-idf: 0.022
Word: 12. Value tf-idf: 0.076
Word: phil. Value tf-idf: 0.102
Word: fraering. Value tf-idf: 0.113
Word: pgf. Value tf-idf: 0.114
Word: srl03. Value tf-idf: 0.121
Word: cacs. Value tf-idf: 0.114
Word: usl. Value tf-idf: 0.112
Word: finally. Value tf-idf: 0.097
Word: this. Value tf-idf: 0.031
Word: isn. Value tf-idf: 0.073
Word: bronze. Value tf-idf: 0.144
Word: age. Value tf-idf: 0.092
Word: please. Value tf-idf: 0.071
Word: try. Value tf-idf: 0.073
Word: to. Value tf-idf: 0.023
Word: remember. Value tf-idf: 0.077
Word: that. Value tf-idf: 0.027
Word: there. Value tf-idf: 0.039
Word: are. Value tf-idf: 0.035
Word: more. Value tf-idf: 0.046
Word: human. Value tf-idf: 0.084
Word: activities. Value tf-idf: 0.091
Word: than. Value tf-idf: 0.052
Word: those. Value tf-idf: 0.060
Word: practiced. Value tf-idf: 0.132
Word: warrior. Value tf-idf: 0.144
Word: caste. Value tf-idf: 0.433
Word: farming. Value tf-idf: 0.144

```
Word: and. Value tf-idf: 0.049
Word: priesthood. Value tf-idf: 0.144
Word: right. Value tf-idf: 0.068
Word: profiting. Value tf-idf: 0.150
Word: is. Value tf-idf: 0.026
Word: blessed. Value tf-idf: 0.150
Word: god. Value tf-idf: 0.119
Word: may. Value tf-idf: 0.054
Word: freely. Value tf-idf: 0.119
Word: blare. Value tf-idf: 0.150
Word: its. Value tf-idf: 0.061
Word: presence. Value tf-idf: 0.119
Word: in. Value tf-idf: 0.024
Word: evening. Value tf-idf: 0.113
Word: twilight. Value tf-idf: 0.139
Word: fred. Value tf-idf: 0.089
```

```
In [29]: ## TFIDF algoritm  for n-grams
         from sklearn.feature_extraction.text import TfidfVectorizer
         tfidf_vect = TfidfVectorizer(ngram_range=(1,3))
         word_tfidf = tfidf_vect.fit_transform(twenty_sci_news.data)
         word_list = tfidf_vect.get_feature_names()
         for n in word_tfidf[0].indices:
             print("Words: %s. Value tf-idf: %0.3f" % (word_list[n], word_tfidf[0,n]))
```

```
Word: from. Value tf-idf: 0.021
Word: flb. Value tf-idf: 0.130
Word: optiplan. Value tf-idf: 0.065
Word: fi. Value tf-idf: 0.054
Word: baube. Value tf-idf: 0.130
Word: tm. Value tf-idf: 0.107
Word: subject. Value tf-idf: 0.011
Word: vandalizing. Value tf-idf: 0.051
Word: the. Value tf-idf: 0.077
Word: sky. Value tf-idf: 0.045
Word: added. Value tf-idf: 0.043
Word: forwarded. Value tf-idf: 0.047
Word: by. Value tf-idf: 0.059
Word: space. Value tf-idf: 0.048
Word: digest. Value tf-idf: 0.046
Word: organization. Value tf-idf: 0.011
Word: via. Value tf-idf: 0.041
Word: international. Value tf-idf: 0.040
Word: university. Value tf-idf: 0.022
Word: original. Value tf-idf: 0.042
Word: sender. Value tf-idf: 0.046
Word: isu. Value tf-idf: 0.049
```

```
Word: vacation. Value tf-idf: 0.049
Word: venari. Value tf-idf: 0.051
Word: cs. Value tf-idf: 0.027
Word: cmu. Value tf-idf: 0.040
Word: edu. Value tf-idf: 0.029
Word: distribution. Value tf-idf: 0.026
Word: sci. Value tf-idf: 0.033
Word: lines. Value tf-idf: 0.011
Word: 12. Value tf-idf: 0.037
Word: phil. Value tf-idf: 0.050
Word: fraering. Value tf-idf: 0.055
Word: pgf. Value tf-idf: 0.056
Word: srl03. Value tf-idf: 0.059
Word: cacs. Value tf-idf: 0.056
Word: usl. Value tf-idf: 0.055
Word: finally. Value tf-idf: 0.048
Word: this. Value tf-idf: 0.015
Word: isn. Value tf-idf: 0.036
Word: bronze. Value tf-idf: 0.071
Word: age. Value tf-idf: 0.045
Word: please. Value tf-idf: 0.035
Word: try. Value tf-idf: 0.036
Word: to. Value tf-idf: 0.011
Word: remember. Value tf-idf: 0.038
Word: that. Value tf-idf: 0.013
Word: there. Value tf-idf: 0.019
Word: are. Value tf-idf: 0.017
Word: more. Value tf-idf: 0.023
Word: human. Value tf-idf: 0.041
Word: activities. Value tf-idf: 0.044
Word: than. Value tf-idf: 0.025
Word: those. Value tf-idf: 0.030
Word: practiced. Value tf-idf: 0.065
Word: warrior. Value tf-idf: 0.071
Word: caste. Value tf-idf: 0.212
Word: farming. Value tf-idf: 0.071
Word: and. Value tf-idf: 0.024
Word: priesthood. Value tf-idf: 0.071
Word: right. Value tf-idf: 0.033
Word: profiting. Value tf-idf: 0.074
Word: is. Value tf-idf: 0.013
Word: blessed. Value tf-idf: 0.074
Word: god. Value tf-idf: 0.058
Word: may. Value tf-idf: 0.027
Word: freely. Value tf-idf: 0.058
Word: blare. Value tf-idf: 0.074
Word: its. Value tf-idf: 0.030
Word: presence. Value tf-idf: 0.058
```

Word: in. Value tf-idf: 0.012
Word: evening. Value tf-idf: 0.055
Word: twilight. Value tf-idf: 0.068
Word: fred. Value tf-idf: 0.044
Word: from flb. Value tf-idf: 0.071
Word: flb flb. Value tf-idf: 0.065
Word: flb optiplan. Value tf-idf: 0.065
Word: optiplan fi. Value tf-idf: 0.065
Word: fi baube. Value tf-idf: 0.065
Word: baube tm. Value tf-idf: 0.130
Word: tm subject. Value tf-idf: 0.071
Word: subject vandalizing. Value tf-idf: 0.071
Word: vandalizing the. Value tf-idf: 0.051
Word: the sky. Value tf-idf: 0.048
Word: sky added. Value tf-idf: 0.078
Word: added forwarded. Value tf-idf: 0.051
Word: forwarded by. Value tf-idf: 0.051
Word: by space. Value tf-idf: 0.050
Word: space digest. Value tf-idf: 0.049
Word: digest organization. Value tf-idf: 0.051
Word: organization via. Value tf-idf: 0.051
Word: via international. Value tf-idf: 0.051
Word: international space. Value tf-idf: 0.049
Word: space university. Value tf-idf: 0.050
Word: university original. Value tf-idf: 0.051
Word: original sender. Value tf-idf: 0.051
Word: sender isu. Value tf-idf: 0.051
Word: isu vacation. Value tf-idf: 0.051
Word: vacation venari. Value tf-idf: 0.051
Word: venari cs. Value tf-idf: 0.051
Word: cs cmu. Value tf-idf: 0.043
Word: cmu edu. Value tf-idf: 0.040
Word: edu distribution. Value tf-idf: 0.049
Word: distribution sci. Value tf-idf: 0.043
Word: sci lines. Value tf-idf: 0.046
Word: lines 12. Value tf-idf: 0.048
Word: 12 from. Value tf-idf: 0.078
Word: from phil. Value tf-idf: 0.074
Word: phil fraering. Value tf-idf: 0.055
Word: fraering pgf. Value tf-idf: 0.074
Word: pgf srl03. Value tf-idf: 0.060
Word: srl03 cacs. Value tf-idf: 0.059
Word: cacs usl. Value tf-idf: 0.056
Word: usl edu. Value tf-idf: 0.055
Word: edu finally. Value tf-idf: 0.074
Word: finally this. Value tf-idf: 0.071
Word: this isn. Value tf-idf: 0.053
Word: isn the. Value tf-idf: 0.062

Word: the bronze. Value tf-idf: 0.071
Word: bronze age. Value tf-idf: 0.071
Word: age please. Value tf-idf: 0.074
Word: please try. Value tf-idf: 0.066
Word: try to. Value tf-idf: 0.044
Word: to remember. Value tf-idf: 0.058
Word: remember that. Value tf-idf: 0.054
Word: that there. Value tf-idf: 0.041
Word: there are. Value tf-idf: 0.034
Word: are more. Value tf-idf: 0.057
Word: more human. Value tf-idf: 0.071
Word: human activities. Value tf-idf: 0.071
Word: activities than. Value tf-idf: 0.068
Word: than those. Value tf-idf: 0.066
Word: those practiced. Value tf-idf: 0.071
Word: practiced by. Value tf-idf: 0.071
Word: by the. Value tf-idf: 0.032
Word: the warrior. Value tf-idf: 0.071
Word: warrior caste. Value tf-idf: 0.071
Word: caste the. Value tf-idf: 0.071
Word: the farming. Value tf-idf: 0.071
Word: farming caste. Value tf-idf: 0.071
Word: caste and. Value tf-idf: 0.071
Word: and the. Value tf-idf: 0.027
Word: the priesthood. Value tf-idf: 0.071
Word: priesthood right. Value tf-idf: 0.074
Word: right the. Value tf-idf: 0.068
Word: the profiting. Value tf-idf: 0.074
Word: profiting caste. Value tf-idf: 0.074
Word: caste is. Value tf-idf: 0.074
Word: is blessed. Value tf-idf: 0.074
Word: blessed by. Value tf-idf: 0.074
Word: by god. Value tf-idf: 0.071
Word: god and. Value tf-idf: 0.074
Word: and may. Value tf-idf: 0.055
Word: may freely. Value tf-idf: 0.074
Word: freely blare. Value tf-idf: 0.074
Word: blare its. Value tf-idf: 0.074
Word: its presence. Value tf-idf: 0.074
Word: presence in. Value tf-idf: 0.071
Word: in the. Value tf-idf: 0.019
Word: the evening. Value tf-idf: 0.066
Word: evening twilight. Value tf-idf: 0.074
Word: twilight fred. Value tf-idf: 0.078
Word: fred baube. Value tf-idf: 0.071
Word: from flb flb. Value tf-idf: 0.071
Word: flb flb optiplan. Value tf-idf: 0.065
Word: flb optiplan fi. Value tf-idf: 0.065

Word: optiplan fi baube. Value tf-idf: 0.065
Word: fi baube tm. Value tf-idf: 0.065
Word: baube tm subject. Value tf-idf: 0.071
Word: tm subject vandalizing. Value tf-idf: 0.078
Word: subject vandalizing the. Value tf-idf: 0.071
Word: vandalizing the sky. Value tf-idf: 0.051
Word: the sky added. Value tf-idf: 0.078
Word: sky added forwarded. Value tf-idf: 0.078
Word: added forwarded by. Value tf-idf: 0.051
Word: forwarded by space. Value tf-idf: 0.051
Word: by space digest. Value tf-idf: 0.051
Word: space digest organization. Value tf-idf: 0.051
Word: digest organization via. Value tf-idf: 0.051
Word: organization via international. Value tf-idf: 0.051
Word: via international space. Value tf-idf: 0.051
Word: international space university. Value tf-idf: 0.050
Word: space university original. Value tf-idf: 0.051
Word: university original sender. Value tf-idf: 0.051
Word: original sender isu. Value tf-idf: 0.051
Word: sender isu vacation. Value tf-idf: 0.051
Word: isu vacation venari. Value tf-idf: 0.051
Word: vacation venari cs. Value tf-idf: 0.051
Word: venari cs cmu. Value tf-idf: 0.051
Word: cs cmu edu. Value tf-idf: 0.043
Word: cmu edu distribution. Value tf-idf: 0.051
Word: edu distribution sci. Value tf-idf: 0.050
Word: distribution sci lines. Value tf-idf: 0.047
Word: sci lines 12. Value tf-idf: 0.074
Word: lines 12 from. Value tf-idf: 0.078
Word: 12 from phil. Value tf-idf: 0.078
Word: from phil fraering. Value tf-idf: 0.074
Word: phil fraering pgf. Value tf-idf: 0.074
Word: fraering pgf srl03. Value tf-idf: 0.074
Word: pgf srl03 cacs. Value tf-idf: 0.060
Word: srl03 cacs usl. Value tf-idf: 0.059
Word: cacs usl edu. Value tf-idf: 0.056
Word: usl edu finally. Value tf-idf: 0.074
Word: edu finally this. Value tf-idf: 0.074
Word: finally this isn. Value tf-idf: 0.071
Word: this isn the. Value tf-idf: 0.065
Word: isn the bronze. Value tf-idf: 0.071
Word: the bronze age. Value tf-idf: 0.071
Word: bronze age please. Value tf-idf: 0.074
Word: age please try. Value tf-idf: 0.074
Word: please try to. Value tf-idf: 0.068
Word: try to remember. Value tf-idf: 0.071
Word: to remember that. Value tf-idf: 0.062
Word: remember that there. Value tf-idf: 0.071

```
Word: that there are. Value tf-idf: 0.056
Word: there are more. Value tf-idf: 0.065
Word: are more human. Value tf-idf: 0.071
Word: more human activities. Value tf-idf: 0.071
Word: human activities than. Value tf-idf: 0.071
Word: activities than those. Value tf-idf: 0.071
Word: than those practiced. Value tf-idf: 0.071
Word: those practiced by. Value tf-idf: 0.071
Word: practiced by the. Value tf-idf: 0.071
Word: by the warrior. Value tf-idf: 0.071
Word: the warrior caste. Value tf-idf: 0.071
Word: warrior caste the. Value tf-idf: 0.071
Word: caste the farming. Value tf-idf: 0.071
Word: the farming caste. Value tf-idf: 0.071
Word: farming caste and. Value tf-idf: 0.071
Word: caste and the. Value tf-idf: 0.071
Word: and the priesthood. Value tf-idf: 0.071
Word: the priesthood right. Value tf-idf: 0.074
Word: priesthood right the. Value tf-idf: 0.074
Word: right the profiting. Value tf-idf: 0.074
Word: the profiting caste. Value tf-idf: 0.074
Word: profiting caste is. Value tf-idf: 0.074
Word: caste is blessed. Value tf-idf: 0.074
Word: is blessed by. Value tf-idf: 0.074
Word: blessed by god. Value tf-idf: 0.074
Word: by god and. Value tf-idf: 0.074
Word: god and may. Value tf-idf: 0.074
Word: and may freely. Value tf-idf: 0.074
Word: may freely blare. Value tf-idf: 0.074
Word: freely blare its. Value tf-idf: 0.074
Word: blare its presence. Value tf-idf: 0.074
Word: its presence in. Value tf-idf: 0.074
Word: presence in the. Value tf-idf: 0.074
Word: in the evening. Value tf-idf: 0.068
Word: the evening twilight. Value tf-idf: 0.074
Word: evening twilight fred. Value tf-idf: 0.078
Word: twilight fred baube. Value tf-idf: 0.078
Word: fred baube tm. Value tf-idf: 0.071
```

In [ ]: