In [518]:
```
"""Ankita Biswas (ab8ky@virginia.edu)
DS5001
06 May 2022"""
```

Out[518]:
```
'Ankita Biswas (ab8ky@virginia.edu)\nDS5001\n06 May 2022'
```

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import nltk
import re
import matplotlib as mpl
mpl.rcParams['legend.frameon'] = False
mpl.rcParams['font.family'] = 'sans-serif'
mpl.rcParams['font.sans-serif'] = 'Helvetica'
mpl.rcParams['font.size'] = 18
```

In [2]:
```
import numpy as np
import scipy as sp
from sklearn.neighbors import KernelDensity as KDE
from nltk.corpus import stopwords
```

In [255]:
```
def generate_tokens(filename, keep_whitespace = True):
    data = np.load(filename)
    data = data.tolist()
    OHCO = ['para_num', 'sent_num', 'token_num']
    PARAS = pd.DataFrame(data, columns=['para_str'])
    PARAS.index.names=OHCO[:1]
    SENTS = PARAS['para_str'].str.split(r'[.?!;:"]+', expand=True).st
    .to_frame().rename(columns={0:'sent_str'})
    SENTS.index.names = OHCO[:2]
    SENTS.sent_str = SENTS.sent_str.str.strip()
    if keep_whitespace:
        TOKENS = SENTS.sent_str\
                  .apply(lambda x: pd.Series(nltk.pos_tag(nltk.word_to
                  .stack()\
                  .to_frame('pos_tuple')
    else:
        TOKENS = SENTS.sent_str\
                  .apply(lambda x: pd.Series(nltk.pos_tag(nltk.Whitesp
                  .stack()\
                  .to_frame('pos_tuple')
    TOKENS.index.names = OHCO[:3]
    TOKENS['pos'] = TOKENS.pos_tuple.apply(lambda x: x[1])
    TOKENS['token_str'] = TOKENS.pos_tuple.apply(lambda x: x[0])
    TOKENS['term_str'] = TOKENS.token_str.replace( '[^A-Za-z0-9\-]',
    return TOKENS
```

In [256]:
```
TOKENS1 = generate_tokens('1.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [257]:
```
TOKENS2 = generate_tokens('2.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [258]:
```
TOKENS3 = generate_tokens('3.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [259]:
```
TOKENS4 = generate_tokens('4.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [260]:
```
TOKENS5 = generate_tokens('5.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [261]:
```
TOKENS6 = generate_tokens('6.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [262]:
```
TOKENS7 = generate_tokens('7.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [263]:
```
TOKENS8 = generate_tokens('8.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [264]:
```
TOKENS9 = generate_tokens('9.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [265]:
```
TOKENS10 = generate_tokens('10.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [266]:
```
TOKENS11 = generate_tokens('11.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [267]:
```
TOKENS12 = generate_tokens('12.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [268]:
```
TOKENS13 = generate_tokens('13.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [269]:
```
TOKENS14 = generate_tokens('14.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
```

In [270]:
```
TOKENS15 = generate_tokens('15.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [271]:
```
TOKENS16 = generate_tokens('16.npy')
```

```
<ipython-input-255-fa34c4f45527>:13: DeprecationWarning:

The default dtype for empty Series will be 'object' instead of 'flo
at64' in a future version. Specify a dtype explicitly to silence th
is warning.
```

In [272]:
```
TOKENS1['doc_id'] = 1
TOKENS2['doc_id'] = 2
TOKENS3['doc_id'] = 3
TOKENS4['doc_id'] = 4
TOKENS5['doc_id'] = 5
TOKENS6['doc_id'] = 6
TOKENS7['doc_id'] = 7
TOKENS8['doc_id'] = 8
TOKENS9['doc_id'] = 9
TOKENS10['doc_id'] = 10
TOKENS11['doc_id'] = 11
TOKENS12['doc_id'] = 12
TOKENS13['doc_id'] = 13
TOKENS14['doc_id'] = 14
TOKENS15['doc_id'] = 15
#TOKENS16['doc_id'] = 16
```

In [273]:
```
CORPUS = pd.concat([TOKENS1, TOKENS2, TOKENS3, TOKENS4, TOKENS5, TOKE
```

In [274]:
```
CORPUS.reset_index(inplace=True, level=['para_num', 'sent_num', 'toke
```

In [275]:
```
CORPUS
```

Out[275]:

	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str	doc_id
0	0	0	0	(The, DT)	DT	The	the	1
1	0	0	1	(variation, NN)	NN	variation	variation	1
2	0	0	2	(of, IN)	IN	of	of	1
3	0	0	3	(the, DT)	DT	the	the	1
4	0	0	4	(lattice, NN)	NN	lattice	lattice	1
...	...	...	...	...	...	...	...	...
57617	20	0	28	(refractory, JJ)	JJ	refractory	refractory	15

	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str	doc_id
57618	20	0	29	(metal, NN)	NN	metal	metal	15
57619	20	0	30	(in, IN)	IN	in	in	15
57620	20	0	31	(the, DT)	DT	the	the	15
57621	20	0	32	(composition, NN)	NN	composition	composition	15

In [276]:
```
OHCO = ['para_num', 'sent_num', 'token_num']
OHCO2 = ['doc_id'] + OHCO
CORPUS = CORPUS.set_index(OHCO2)
```

In [277]:
```
CORPUS
```

Out[277]:

doc_id	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str
			0	(The, DT)	DT	The	the
			1	(variation, NN)	NN	variation	variation
1	0	0	2	(of, IN)	IN	of	of
			3	(the, DT)	DT	the	the
			4	(lattice, NN)	NN	lattice	lattice
...	...	...	...	...	...	...	...
			28	(refractory, JJ)	JJ	refractory	refractory
			29	(metal, NN)	NN	metal	metal
15	20	0	30	(in, IN)	IN	in	in
			31	(the, DT)	DT	the	the
			32	(composition, NN)	NN	composition	composition

57622 rows Ч 4 columns

In [278]:
```
CORPUS['term_str'] = CORPUS['term_str'].map(lambda x: re.sub('[^\w\s
```

In [279]:
```
CORPUS[CORPUS.term_str == ''].token_str.value_counts()
```

Out[279]:

```
,       2920
)        948
(        931
=        193
%        160
]        130
*        121
[        105
’         76
”         76
_         66
<         48
{         46
—         41
«         39
+         38
'         37
|         32
°         31
~         30
>         28
“         26
}         19
@         16
&         14
-         11
®         11
‘          9
/          9
'°         8
€          6
'®         6
¢          6
$          6
#          4
£          4
§          3
--         3
»          2
°°         2
é          2
''         2
—~         2
```

In [280]:
```
CORPUS = CORPUS[CORPUS.term_str != '']
```

In [281]:
```
CORPUS
```

Out[281]:

doc_id	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str
1	0	0	0	(The, DT)	DT	The	the
			1	(variation, NN)	NN	variation	variation
			2	(of, IN)	IN	of	of
			3	(the, DT)	DT	the	the
			4	(lattice, NN)	NN	lattice	lattice

				pos_tuple	pos	token_str	term_str
doc_id	para_num	sent_num	token_num				
...	...	...	...	...	...	...	...
			28	(refractory, JJ)	JJ	refractory	refractory
			29	(metal, NN)	NN	metal	metal
15	20	0	30	(in, IN)	IN	in	in
			31	(the, DT)	DT	the	the

In [282]:
```
CORPUS.term_str = CORPUS.term_str.str.replace('\d+', '', regex=True)
```

```
/home/digifort/anaconda3/lib/python3.8/site-packages/pandas/core/ge
neric.py:5494: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pan
das-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-
copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/index
ing.html#returning-a-view-versus-a-copy)
```

In [283]:
```
CORPUS
```

Out[283]:

				pos_tuple	pos	token_str	term_str
doc_id	para_num	sent_num	token_num				
			0	(The, DT)	DT	The	the
			1	(variation, NN)	NN	variation	variation
1	0	0	2	(of, IN)	IN	of	of
			3	(the, DT)	DT	the	the
			4	(lattice, NN)	NN	lattice	lattice
...	...	...	...	...	...	...	...
			28	(refractory, JJ)	JJ	refractory	refractory
			29	(metal, NN)	NN	metal	metal
15	20	0	30	(in, IN)	IN	in	in
			31	(the, DT)	DT	the	the
			32	(composition, NN)	NN	composition	composition

51336 rows Ч 4 columns

Removing stopwords from CORPUS

In [284]:
```
nltk.download('stopwords')
stopwords = nltk.corpus.stopwords.words('english')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/digifort/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [285]:
```
CORPUS['bool'] = CORPUS['term_str'].apply(lambda x: 'NaN' if x in sto
```

```
<ipython-input-285-5f24f4292f2d>:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
```

In [286]:
```
CORPUS
```

Out[286]:

				pos_tuple	pos	token_str	term_str	bool
doc_id	para_num	sent_num	token_num					
1	0	0	0	(The, DT)	DT	The	the	NaN
			1	(variation, NN)	NN	variation	variation	1
			2	(of, IN)	IN	of	of	NaN
			3	(the, DT)	DT	the	the	NaN
			4	(lattice, NN)	NN	lattice	lattice	1
...	...	...	...	...	...	...	...	...
			28	(refractory, JJ)	JJ	refractory	refractory	1
			29	(metal, NN)	NN	metal	metal	1
15	20	0	30	(in, IN)	IN	in	in	NaN
			31	(the, DT)	DT	the	the	NaN
			32	(composition, NN)	NN	composition	composition	1

51336 rows Ч 5 columns

In [287]:
```
CORPUS = CORPUS[CORPUS['bool'] != 'NaN']
```

In [288]:
```
CORPUS = CORPUS.dropna(subset=['term_str'])
```

In [289]:
```
CORPUS.drop('bool', axis=1, inplace=True)
```

In [290]:
```
CORPUS
```

Out[290]:

				pos_tuple	pos	token_str	term_str

doc_id	para_num	sent_num	token_num				
			1	(variation, NN)	NN	variation	variation
			4	(lattice, NN)	NN	lattice	lattice
1	0	0	5	(constant, NN)	NN	constant	constant
			7	(alloying, VBG)	VBG	alloying	alloying
			8	(elements, NNS)	NNS	elements	elements
...	...	...	...	...	...	...	...
			22	(based, VBN)	VBN	based	based
			25	(role, NN)	NN	role	role
15	20	0	28	(refractory, JJ)	JJ	refractory	refractory
			29	(metal, NN)	NN	metal	metal
			32	(composition, NN)	NN	composition	composition

In [39]:
```
CORPUS.to_csv('CORPUS.csv')
```

Time Token Matrix

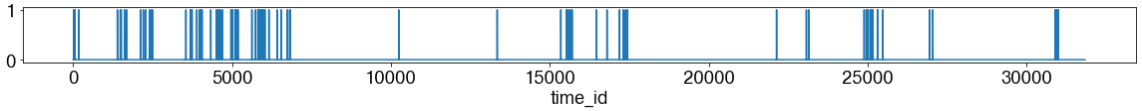
In [40]:
```
TTM = pd.get_dummies(CORPUS['term_str'],
                     columns=['term_str'],
                     prefix='',
                     prefix_sep='',
                     drop_first=True)\
    .reset_index(drop=True).iloc[:,1:]
TTM.index.name = 'time_id'
```

In [41]:
```
import matplotlib.pyplot as plt
```

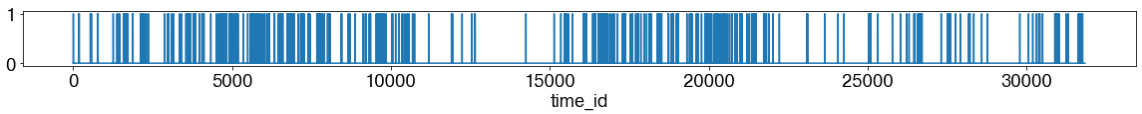
In [42]:
```
cfg = {'figsize': (20,1)}
```

In [43]:
```
TTM['lattice'].plot(**cfg);
```

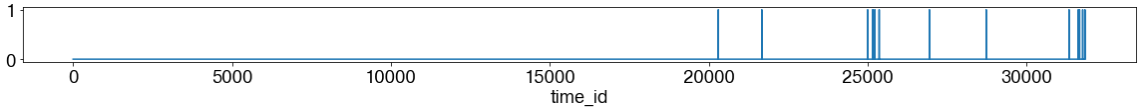
In [44]:
```
TTM['entropy'].plot(**cfg);
```



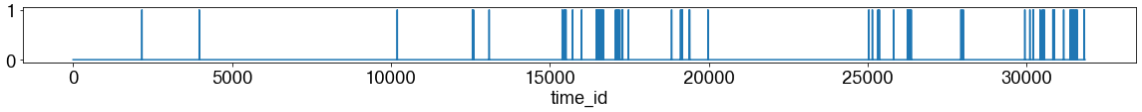
In [45]:
```
TTM['alloy'].plot(**cfg);
```



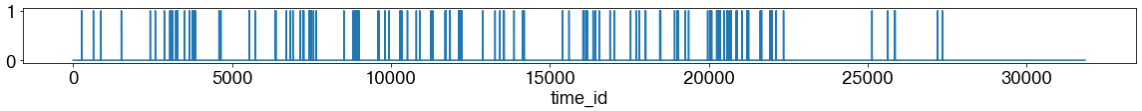
In [46]:
```
TTM['refractory'].plot(**cfg);
```



In [47]:
```
TTM['hardness'].plot(**cfg);
```

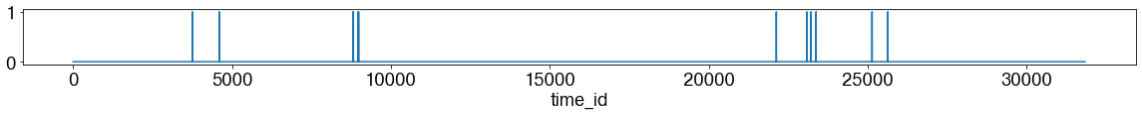


In [48]:
```
TTM['fcc'].plot(**cfg);
```



In [49]:
```
TTM['bcc'].plot(**cfg);
```

In [50]:
```
TTM['hcp'].plot(**cfg);
```



In [51]:
```
R = CORPUS['term_str'].reset_index(drop=True).to_frame().reset_index(
```

In [52]:
```
R
```

Out[52]:

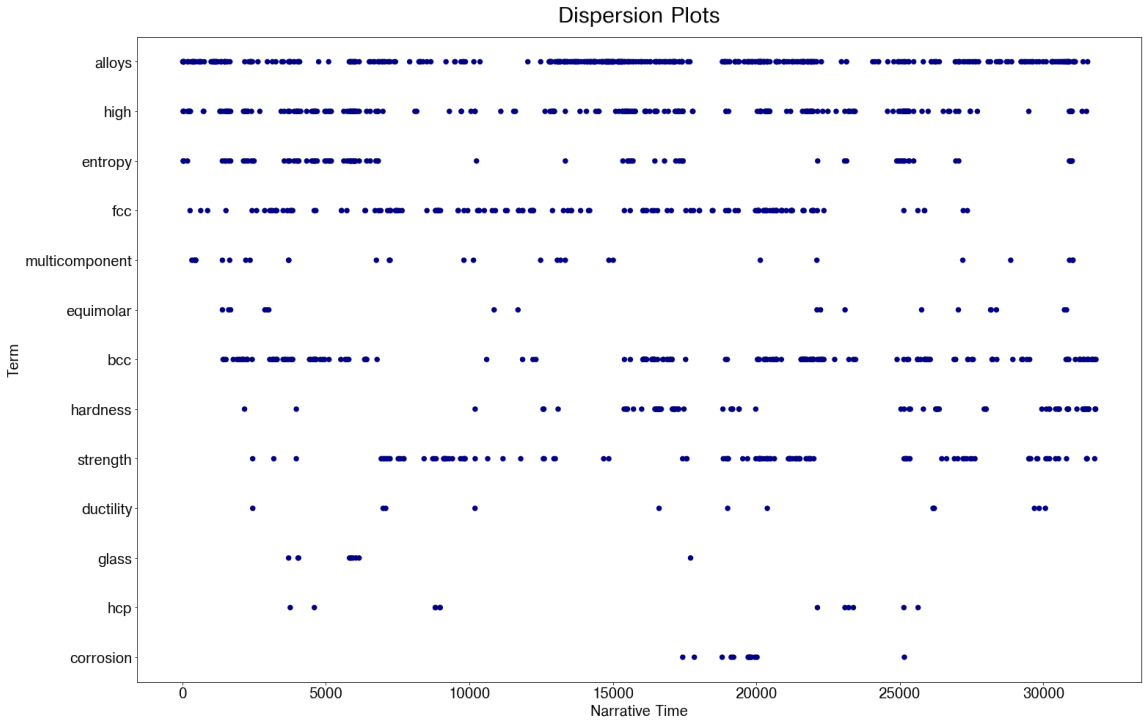
	offset	term_str
0	0	variation
1	1	lattice
2	2	constant
3	3	alloying
4	4	elements
...	...	...
31836	31836	based
31837	31837	role
31838	31838	refractory
31839	31839	metal
31840	31840	composition

31841 rows Ч 2 columns

In [53]:
```
names = ['bcc', 'fcc', 'hcp', 'high', 'strength', 'ductility', 'entro
```

In [54]:
```
X = B[B.term_str.isin(names)]
plt.figure(figsize=(22, len(names)))
ax = sns.stripplot(y='term_str', x='offset', data=X, orient='h', mark
ax.set_title('Dispersion Plots', size=30, pad=20)
ax.set_xlabel('Narrative Time', size=20)
ax.set_ylabel('Term', size=20)

plt.xticks(rotation=0, fontsize=20)
plt.yticks(rotation=0, fontsize=20)
plt.tight_layout()
#plt.show()
plt.savefig('dispersion_plot.png', dpi=300)
```

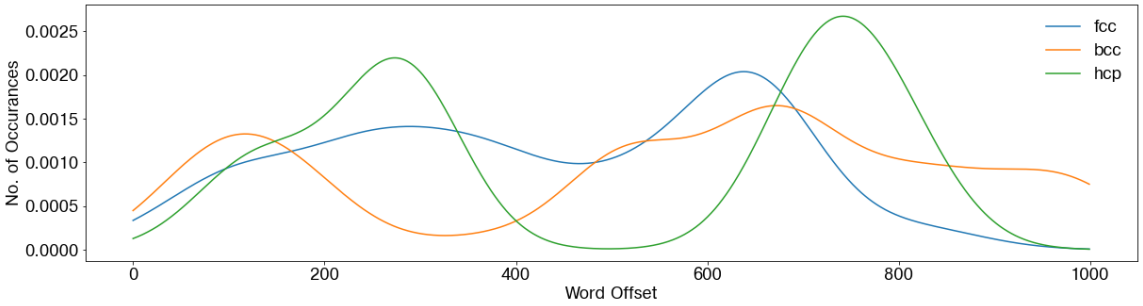


In [55]:
```
kde_kernel = 'gaussian'
kde_bandwidth = 2000
kde_samples = 1000
```

In [56]:
```
X = B.reset_index().groupby(['term_str']).offset.apply(lambda x: x.to
X['x'] = X.apply(lambda x: np.array(x.offset)[:, np.newaxis], 1)
scale_max = B.offset.max() # THIS IS CRUCIAL
x_axis = np.linspace(0, scale_max, kde_samples)[:, np.newaxis]
X['kde'] = X.apply(lambda row: KDE(kernel=kde_kernel, bandwidth=kde_b
X['scores'] = X.apply(lambda row: row.kde.score_samples(x_axis), axis
PLOTS = X.apply(lambda row: pd.Series(np.exp(row.scores) * (scale_ma
```

In [57]:
```
FIG = dict(figsize=(20, 5))
PLOTS.loc['fcc'].plot(**FIG, legend=True);
PLOTS.loc['bcc'].plot(**FIG, legend = True);
PLOTS.loc['hcp'].plot(**FIG, legend = True);
plt.xlabel('Word Offset')
plt.ylabel('No. of Occurances')
```

Out[57]:
```
Text(0, 0.5, 'No. of Occurances')
```



Creation of LIB

In [60]:
```
title = ['Quantitative determination of the lattice constant in high
         'High entropy multicomponent WMoNbZrV alloy processed bymech
         'Mapping the magnetic transition temperatures for medium- ar
        'Mechano-chemical synthesis, thermal stability and phase evol
        'First-principles-based prediction of yield strength in the F
        'Structure and properties of equiatomic CoCrFeNiMn alloy fabr
        'Assessing elastic property and solid-solution strengthening
        'Fast production of high entropy alloys (CoCrFeNiAlxTiy) by e
        'Chemical complexity induced local structural distortion in M
        'Microstructures and properties of Al0.3CoCrFeNiMnx high-entr
        'Vanadium is an optimal element for strengthening in both fcc
        'High-entropy alloy superconductors on an \u03B1-Mn lattice',
        'First-principle calculation investigation of NbMoTaW based r
        'Contribution of Lattice Distortion to Solid Solution Strengt
        'Role of Various Parameters in the Formation of the Physicome
```

In [61]:
```
authors = ['Zhijun Wang, Qingfeng Wu, Wenquan Zhou, Feng He, Chunyan
          'Dariusz Oleszak, Anna Antolak-Dudka, Tadeusz Kulik',
          u'Shuo Huang, Erik Holmström, Olle Eriksson, Levente Vitos'
          'Vikas Shivam, Joysurya Basu, Yagnesh Shadangi, Manish Kuma
          'Binglun Yin, William A. Curtin',
          'A.S. Rogachev, S.G. Vadchenko, N.A. Kochetov, S. Rouvimov,
          'Zhi-biao Yang, Jian Sun', 'Azmi Erdogan, Tuba Yener, Sakir
          'Fuxiang Zhang, Yang Tong, Ke Jin, Hongbin Bei, William J.
          'Sze-Kwan Wong, Tao-Tsung Shun, Chieh-Hsiang Chang, Che-Fu
          'Binglun Yin, Francesco Maresca, W.A. Curtin',
          'Karoline Stolze, F. Alex Cevallos, Tai Kong, Robert J. Cav
          'Y.L. Hu, L.H. Bai, Y.G. Tong, D.Y. Deng, X.B. Liang, J. Zh
          u'H. Chen, A. Kauffmann, S. Laube, I.-C. Choi, R. Schwaiger
          'V. F. Gorban, N. A. Krapivka, S. A. Firstov, D. V. Kuriler
```

In [62]:
```
publisher = ['Scripta Materialia 162 (2019) 468-471', 'Materials Lett
             'Intermetallics 95 (2018) 80-84', 'Journal of Alloys and
            'npj Computational Materials (2019) 5:14', 'Journal of Al
            'Journal of Materials Research volume 33, pages 2763-2774
            'Materials Research Letters, 6:8, 450-455', 'Materials Ch
            'Acta Materialia 188 (2020) 486-491', 'Journal of Materia
```

```
            'Journal of Alloys and Compounds 827 (2020) 153963',
            'Metallurgical and Materials Transactions A volume 49, pa
            'Physics of Metals and Metallography  Volume 119  Issue 5
```

In [63]:
```
LIB = pd.DataFrame(title, columns = ['title'])
```

In [64]:
```
LIB['authors'] = authors
```

In [65]:
```
LIB['publisher'] = publisher
```

In [67]:
```
LIB['doc_id'] = [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]
```

In [68]:
```
LIB.set_index('doc_id')
```

Out[68]:

doc_id	title	authors	publisher
1	Quantitative determination of the lattice cons...	Zhijun Wang, Qingfeng Wu, Wenquan Zhou, Feng H...	Scripta Materialia 162 (2019) 468-471
2	High entropy multicomponent WMoNbZrV alloy pro...	Dariusz Oleszak, Anna Antolak-Dudka, Tadeusz K...	Materials Letters 232 (2018) 160-162
3	Mapping the magnetic transition temperatures f...	Shuo Huang, Erik Holmstrцm, Olle Eriksson, Lev...	Intermetallics 95 (2018) 80-84
4	Mechano-chemical synthesis, thermal stability ...	Vikas Shivam, Joysurya Basu, Yagnesh Shadangi,...	Journal of Alloys and Compounds 757 (2018) 87-97
5	First-principles-based prediction of yield str...	Binglun Yin, William A. Curtin	npj Computational Materials (2019) 5:14
6	Structure and properties of equiatomic CoCrFeN...	A.S. Rogachev, S.G. Vadchenko, N.A. Kochetov, ...	Journal of Alloys and Compounds 805 (2019) 123...
7	Assessing elastic property and solid-solution ...	Zhi-biao Yang, Jian Sun	Journal of Materials Research volume 33, pages...
8	Fast production of high entropy alloys (CoCrFe...	Azmi Erdogan, Tuba Yener, Sakin Zeytin	Vacuum 155 (2018) 64-72
9	Chemical complexity induced local structural d...	Fuxiang Zhang, Yang Tong, Ke Jin, Hongbin Bei,...	Materials Research Letters, 6:8, 450-455
10	Microstructures and properties of Al0.3CoCrFeN...	Sze-Kwan Wong, Tao-Tsung Shun, Chieh-Hsiang Ch...	Materials Chemistry and Physics 210 (2018) 146...
11	Vanadium is an optimal element for strengtheni...	Binglun Yin, Francesco Maresca, W.A. Curtin	Acta Materialia 188 (2020) 486-491
12	High-entropy alloy superconductors on an α-Mn ...	Karoline Stolze, F. Alex Cevallos, Tai Kong, R...	Journal of Materials Chemistry C, 2018, 6, 10441
13	First-principle calculation investigation of N...	Y.L. Hu, L.H. Bai, Y.G. Tong, D.Y. Deng, X.B. ...	Journal of Alloys and Compounds 827 (2020) 153963
14	Contribution of Lattice Distortion to Solid So...	H. Chen, A. Kauffmann, S. Laube, I.-C. Choi, R...	Metallurgical and Materials Transactions A vol...
15	Role of Various Parameters in the Formation of...	V. F. Gorban, N. A. Krapivka, S. A. Firstov, D...	Physics of Metals and Metallography, Volume 11...

```
CORPUS['term_str'] = CORPUS['term_str'].str.replace('[^\w\s]', '', regex=True)

CORPUS['term_str'] = CORPUS['term_str'].str.replace('\d+', '', regex=True,
inplace=True)
```

In [69]:
```
CORPUS
```

Out[69]:

				pos_tuple	pos	token_str	term_str
doc_id	para_num	sent_num	token_num				
			1	(variation, NN)	NN	variation	variation
			4	(lattice, NN)	NN	lattice	lattice
1	0	0	5	(constant, NN)	NN	constant	constant
			7	(alloying, VBG)	VBG	alloying	alloying
			8	(elements, NNS)	NNS	elements	elements
...	...	...	...	...	...	...	...
			22	(based, VBN)	VBN	based	based
			25	(role, NN)	NN	role	role
15	20	0	28	(refractory, JJ)	JJ	refractory	refractory
			29	(metal, NN)	NN	metal	metal
			32	(composition, NN)	NN	composition	composition

31841 rows Ч 4 columns

In [70]:
```
VOCAB = CORPUS.term_str.value_counts().to_frame('n').sort_index()
VOCAB.index.name = 'term_str'
VOCAB['n_chars'] = VOCAB.index.str.len()
VOCAB['p'] = VOCAB.n / VOCAB.n.sum()
VOCAB['i'] = -np.log2(VOCAB.p)
```

In [71]:
```
VOCAB
```

Out[71]:

	n	n_chars	p	i
term_str				
	2580	0	0.081028	3.625443
aa	1	2	0.000031	14.958598
aaa	3	3	0.000094	13.373636
aaaa	2	4	0.000063	13.958598
aaaaaadaaag	1	11	0.000031	14.958598
...	...	...	...	...
zrrez	1	5	0.000031	14.958598
zs	2	2	0.000063	13.958598
ztnb	1	4	0.000031	14.958598
zwick	1	5	0.000031	14.958598
zz	5	2	0.000157	12.636670

4499 rows × 4 columns

In [72]:
```
VOCAB['max_pos'] = CORPUS[['term_str', 'pos']].value_counts().unstack(
```

In [73]:
```
VOCAB
```

Out[73]:

term_str	n	n_chars	p	i	max_pos
	2580	0	0.081028	3.625443	CD
aa	1	2	0.000031	14.958598	JJ
aaa	3	3	0.000094	13.373636	NNP
aaaa	2	4	0.000063	13.958598	NNP
aaaaaadaaag	1	11	0.000031	14.958598	NNP
...	...	...	...	...	...
zrrez	1	5	0.000031	14.958598	NNP
zs	2	2	0.000063	13.958598	NNP
ztnb	1	4	0.000031	14.958598	NNP
zwick	1	5	0.000031	14.958598	NNP
zz	5	2	0.000157	12.636670	NNP

4499 rows × 5 columns

In [74]:
```
VOCAB.to_csv('VOCAB.csv')
```

CORPUS = pd.read_csv('CORPUS.csv')

In [75]:
```
CORPUS
```

Out[75]:

doc_id	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str
			1	(variation, NN)	NN	variation	variation
			4	(lattice, NN)	NN	lattice	lattice
1	0	0	5	(constant, NN)	NN	constant	constant
			7	(alloying, VBG)	VBG	alloying	alloying
			8	(elements, NNS)	NNS	elements	elements
...	...	...	...	...	...	...	...
			22	(based, VBN)	VBN	based	based
			25	(role, NN)	NN	role	role
15	20	0	28	(refractory, JJ)	JJ	refractory	refractory
			29	(metal, NN)	NN	metal	metal
			32	(composition, NN)	NN	composition	composition

31841 rows × 4 columns

In [76]:
```
TPM = CORPUS[['term_str','pos']].value_counts().unstack()
```

In [77]:
```
TPM
```

Out[77]:

pos	$	CC	CD	DT	EX	FW	IN	JJ	JJR	JJS	...	SYM	UH	V
term_str														
	1.0	NaN	2414.0	NaN	NaN	NaN	NaN	77.0	NaN	NaN	...	NaN	NaN	5
aa	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	...	NaN	NaN	Na
aaa	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	Na
aaaa	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	Na
aaaaaadaaag	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	Na
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
zrrez	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	Na
zs	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	Na
ztnb	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	Na
zwick	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	Na
zz	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	Na

4499 rows Ч 32 columns

In [78]:
```
VOCAB['n_pos'] = TPM.count(1)
```

In [79]:
```
VOCAB['cat_pos'] = CORPUS[['term_str','pos']].value_counts().to_frame
    .groupby('term_str').pos.apply(lambda x: set(x))
```

In [80]:
```
VOCAB
```

Out[80]:

n	n_chars	p	i	max_pos	n_pos	cat_pos
term_str						
2580	0	0.081028	3.625443	CD	11	{VBP, NNS, NN, VBZ, CD, NNP, VB, $, JJ, VBD, LS}
aa	1	2	0.000031	14.958598	JJ	1	{JJ}
aaa	3	3	0.000094	13.373636	NNP	2	{VBZ, NNP}
aaaa	2	4	0.000063	13.958598	NNP	1	{NNP}
aaaaaadaaag	1	11	0.000031	14.958598	NNP	1	{NNP}
...	...	...	...	...	...	...	...
zrrez	1	5	0.000031	14.958598	NNP	1	{NNP}
zs	2	2	0.000063	13.958598	NNP	1	{NNP}
ztnb	1	4	0.000031	14.958598	NNP	1	{NNP}
zwick	1	5	0.000031	14.958598	NNP	1	{NNP}
zz	5	2	0.000157	12.636670	NNP	1	{NNP}

4499 rows Ч 7 columns

In [81]:
```
sw = pd.DataFrame(nltk.corpus.stopwords.words('english'), columns=['t
sw = sw.reset_index().set_index('term_str')
sw.columns = ['dummy']
sw.dummy = 1
```

In [82]:
```
VOCAB['stop'] = VOCAB.index.map(sw.dummy)
VOCAB['stop'] = VOCAB['stop'].fillna(0).astype('int')
```

In [83]:
```
from nltk.stem.porter import PorterStemmer
stemmer1 = PorterStemmer()
VOCAB['stem_porter'] = VOCAB.apply(lambda x: stemmer1.stem(x.name), 1

from nltk.stem.snowball import SnowballStemmer
stemmer2 = SnowballStemmer("english")
VOCAB['stem_snowball'] = VOCAB.apply(lambda x: stemmer2.stem(x.name),

from nltk.stem.lancaster import LancasterStemmer
stemmer3 = LancasterStemmer()
VOCAB['stem_lancaster'] = VOCAB.apply(lambda x: stemmer3.stem(x.name)
```

In [84]:
```
VOCAB.sample(10)
```

Out[84]:

term_str	n	n_chars	p	i	max_pos	n_pos	cat_pos	stop
modified	2	8	0.000063	13.958598	JJ	2	{VBN, JJ}	0
grain	21	5	0.000660	10.566281	NN	3	{JJ, VB, NN}	0
cooooooooooooooooooo	1	21	0.000031	14.958598	NN	1	{NN}	0
weak	2	4	0.000063	13.958598	JJ	1	{JJ}	0
account	12	7	0.000377	11.373636	NN	2	{VB, NN}	0
needed	6	6	0.000188	12.373636	VBN	1	{VBN}	0
approximations	1	14	0.000031	14.958598	NNS	1	{NNS}	0
adiabatic	1	9	0.000031	14.958598	JJ	1	{JJ}	0
uy	1	2	0.000031	14.958598	JJ	1	{JJ}	0
seem	3	4	0.000094	13.373636	VBP	2	{VBP, VB}	0

In [85]:
```
most_aggressive_stem = VOCAB.stem_lancaster.value_counts().head(1).ir
```

In [86]:
```
most_aggressive_stem
```

Out[86]:
```
'ind'
```

In [87]:
```
VOCAB.query(f"stem_lancaster == '{most_aggressive_stem}'")
```

Out[87]:

term_str	n	n_chars	p	i	max_pos	n_pos	cat_pos	stop	stem_porter
indent	3	6	0.000094	13.373636	NN	1	{NN}	0	indent
indentation	3	11	0.000094	13.373636	NN	1	{NN}	0	indent

term_str	n	n_chars	p	i	max_pos	n_pos	cat_pos	stop	stem_porter
indentations	1	12	0.000031	14.958598	NNS	1	{NNS}	0	indent
indenter	2	8	0.000063	13.958598	NN	1	{NN}	0	indent
indicate	7	8	0.000220	12.151243	VBP	3	{VBP, VB, NN}	0	indic
indicated	6	9	0.000188	12.373636	VBN	2	{VBN, VBD}	0	indic
indicates	23	9	0.000722	10.435036	VBZ	1	{VBZ}	0	indic
indicating	13	10	0.000408	11.258158	VBG	1	{VBG}	0	indic
indication	2	10	0.000063	13.958598	NN	1	{NN}	0	indic

N-gram models

In [88]:
```
import sys
sys.path.append('/home/digifort/Documents/Sem_3/Exploratory_text_anal
```

In [89]:
```
from langmod import NgramCounter, NgramLanguageModel
```

In [90]:
```
files = ['1.npy', '2.npy', '3.npy', '4.npy', '5.npy', '6.npy', '7.npy
        '9.npy', '10.npy', '11.npy', '12.npy', '13.npy', '14.npy', '1
SENTS_final = []
for i in files:
    data = np.load(i)
    data = data.tolist()
    OHCO = ['para_num', 'sent_num', 'token_num']
    PARAS = pd.DataFrame(data, columns=['para_str'])
    PARAS.index.names=OHCO[:1]
    SENTS = PARAS['para_str'].str.split(r'[.?!;:"]+', expand=True).st
           .to_frame().rename(columns={0:'sent_str'})
    SENTS.index.names = OHCO[:2]
    SENTS.sent_str = SENTS.sent_str.str.strip()
    l = SENTS.sent_str.to_list()
    SENTS_final.append(l)
```

In [91]:
```
SENTS_final
```

Out[91]:

```
[['The variation of the lattice constant with alloying elements is
an essential issue in alloy design'
```

In [92]:
```
from itertools import chain
```

In [93]:
```
SENTS = list(chain.from_iterable(SENTS_final))
```

In [94]:
```
SENTS
```

Out[94]:
```
['The variation of the lattice constant with alloying elements is a
n essential issue in alloy design',
 'In the traditional single-based alloys, there is a tremendous dat
abase for predicting the lattice constant',
 'However, the traditional database is not suitable to the emerging
multi-principal components alloys referred as high entropy alloys',
 'Here, a framework is proposed to describe the variation of lattic
e constants in high entropy alloys',
 'Based on the quantitative measurement of the lattice constants of
fourteen alloys, we constructed the lattice constant database for c
oncentrated CoCrFeNi alloys',
 'The discrepancy between binary alloys and high entropy alloys rev
ealed the atomic chemical interactions',
 '',
 'As one of the most important physical parameters in crystalline m
aterials, the lattice constant has been involved everywhere in the
material science and engineering',
 'Simply, the thermal expansion can be exactly revealed from the te
mperature-dependent lattice constants Another case is the lattice d
```

In [95]:
```
VOCAB= pd.read_csv('VOCAB.csv')
```

In [96]:
```
VOCAB.set_index('term_str', inplace = True)
```

In [97]:
```
vocab = VOCAB.index.to_list()
```

In [98]:
```
vocab
```

Out[98]:
```
[nan,
 'aa',
 'aaa',
 'aaaa',
 'aaaaaadaaag',
 'aaanaaaag',
 'aaax',
 'aad',
 'aae',
 'aamnns',
 'aas',
 'aat',
 'ab',
 'abcdef',
 'aberration',
 'abilities',
 'ability',
 'abinitial',
 'able',
```

In [99]:
```
train = NgramCounter(SENTS, vocab)
```

In [100]:
```
train.generate()
```

In [101]:
```
model = NgramLanguageModel(train)
model.apply_smoothing()
```

In [102]:
```
test_sents = """
A general theory has been developed to predict the temperature and st
The theory envisions the HEA as an “effective-medium matrix”, and eac
With an additional assumption that the solute/dislocation interactior
""".split('\n')[1:-1]
```

In [103]:
```
test_sents = [s.lower() for s in test_sents]
```

In [104]:
```
test = NgramCounter(test_sents, vocab)
test.generate()
```

In [105]:
```
model.predict(test)
```

In [106]:
```
model.T.S
```

Out[106]:

	sent_str	len	ng_1_ll	pp1	ng_2_ll	pp2	ng_3_ll	
0	a general theory has been developed to predict...	22	-125.464065	52.089188	-139.692371	81.552277	-407.203521	373112.43
1	the theory envisions the hea as an “effective-...	27	-116.672000	19.989704	-157.038037	56.345309	-474.988337	197591.39
2	with an additional assumption that the solute/...	40	-208.993597	37.396722	-254.140538	81.770777	-696.185291	173506.75

In [107]:
```
model.generate_text()
```

```
01. <UNK> EMAIL.

02. <UNK> CASE <UNK> <UNK> <UNK> <UNK> FCC LATTICE SITES.

03. <UNK> <UNK> MM DIAMETER <UNK> <UNK> FUNCTION <UNK> DECREASING T
EMPERATURE.

04. <UNK> MAIN ALLOYING ELEMENTS.

05. <UNK>.

06. <UNK> <UNK> SOLID SOLUTION REMAINS NEARLY <UNK> ENHANCEMENT <UN
K> <UNK> RAPID <UNK> CHEMICAL INTERACTIONS.

07. <UNK>.

08. <UNK> <UNK> LATTICE CONSTANT <UNK> ELASTIC CONSTANTS WOULD BRIN
G <UNK> PREDICTIONS <UNK> <UNK> <UNK>.

09. <UNK> SINGLE EXPERIMENTAL VALUE <UNK> <UNK> CONVENTIONAL ONES B
ASED <UNK> CLOSEST PACKING RATHER <UNK> INTERMETALLIC PHASES <UNK>
IDENTIFIED BASED <UNK> <UNK> <UNK> <UNK> MODULUS <UNK> <UNK>.

10. <UNK>.

11. <UNK> METHODS DESCRIBED <UNK> <UNK> <UNK>.

12. <UNK> LEAST FIVE PRINCIPAL ELEMENTS <UNK> <UNK> COMPONENTS <UNK
> COMPOSITION <UNK> <UNK>.

13. <UNK> <UNK> RANDOM ALLOYS <UNK> <UNK> CLOSE LATTICE <UNK> SHEAR
```

In [108]:
```
V = len(vocab)
R = []
for i in range(3):
    N = V**(i+1)
    H = (train.LM[i]['mle'] * np.log2(1/train.LM[i]['mle'])).sum()
    Hmax = np.log2(N)
    R.append(int(round((1 - H/Hmax)**2) * 100))
```

In [109]:
```
R
```

Out[109]:
```
[59, 66, 71]
```

In [110]:
```
LTR.set_index('doc_id', inplace=True)
```

In [111]:
```
LTR
```

Out[111]:

doc_id	title	authors	publisher
1	Quantitative determination of the lattice cons...	Zhijun Wang, Qingfeng Wu, Wenquan Zhou, Feng H...	Scripta Materialia 162 (2019) 468-471
2	High entropy multicomponent WMoNbZrV alloy pro...	Dariusz Oleszak, Anna Antolak-Dudka, Tadeusz K...	Materials Letters 232 (2018) 160-162
3	Mapping the magnetic transition temperatures f...	Shuo Huang, Erik Holmstrцm, Olle Eriksson, Lev...	Intermetallics 95 (2018) 80-84

doc_id	title	authors	publisher
4	Mechano-chemical synthesis, thermal stability ...	Vikas Shivam, Joysurya Basu, Yagnesh Shadangi,...	Journal of Alloys and Compounds 757 (2018) 87-97
5	First-principles-based prediction of yield str...	Binglun Yin, William A. Curtin	npj Computational Materials (2019) 5:14
6	Structure and properties of equiatomic CoCrFeN...	A.S. Rogachev, S.G. Vadchenko, N.A. Kochetov, ...	Journal of Alloys and Compounds 805 (2019) 123...
7	Assessing elastic property and solid-solution ...	Zhi-biao Yang, Jian Sun	Journal of Materials Research volume 33, pages...
8	Fast production of high entropy alloys (CoCrFe...	Azmi Erdogan, Tuba Yener, Sakin Zeytin	Vacuum 155 (2018) 64-72
9	Chemical complexity induced local structural d...	Fuxiang Zhang, Yang Tong, Ke Jin, Hongbin Bei,...	Materials Research Letters, 6:8, 450-455
10	Microstructures and properties of Al0.3CoCrFeN...	Sze-Kwan Wong, Tao-Tsung Shun, Chieh-Hsiang Ch...	Materials Chemistry and Physics 210 (2018) 146...
11	Vanadium is an optimal element for strengtheni...	Binglun Yin, Francesco Maresca, W.A. Curtin	Acta Materialia 188 (2020) 486-491
12	High-entropy alloy superconductors on an α-Mn ...	Karoline Stolze, F. Alex Cevallos, Tai Kong, R...	Journal of Materials Chemistry C, 2018, 6, 10441
13	First-principle calculation investigation of N...	Y.L. Hu, L.H. Bai, Y.G. Tong, D.Y. Deng, X.B. ...	Journal of Alloys and Compounds 827 (2020) 153963
14	Contribution of Lattice	H. Chen, A. Kauffmann, S.	Metallurgical and Materials

Calculation of TFIDF

CORPUS['term_str'] = CORPUS['token']

In [112]:
```
VOCAB_new = CORPUS.term_str.value_counts().to_frame('n')
VOCAB_new.index.name = 'term_str'
VOCAB_new['p'] = VOCAB_new.n / VOCAB_new.n.sum()
VOCAB_new['i'] = np.log2(1/VOCAB_new.p)
VOCAB_new['max_pos'] = CORPUS.reset_index().value_counts(['term_str'
```

In [113]:
```
VOCAB_new
```

Out[113]:

term_str	n	p	i	max_pos
	2580	0.081028	3.625443	CD
alloys	481	0.015106	6.048705	NNS
alloy	376	0.011809	6.404009	NN
lattice	367	0.011526	6.438962	NN
phase	270	0.008480	6.881782	NN
...	...	...	...	...

	n	p	i	max_pos
term_str				
qcaie	1	0.000031	14.958598	NNP
restricted	1	0.000031	14.958598	VBN
mixtures	1	0.000031	14.958598	NNS
dc	1	0.000031	14.958598	NNP

In [114]:
```
def create_bow(CORPUS, bag, item_type='term_str'):
    BOW = CORPUS.groupby(bag+[item_type])[item_type].count().to_frame
    return BOW
```

In [115]:
```
def get_tfidf(BOW, tf_method='max', df_method='standard', item_type='

    DTCM = BOW.n.unstack() # Create Doc-Term Count Matrix

    if tf_method == 'sum':
        TF = (DTCM.T / DTCM.T.sum()).T
    elif tf_method == 'max':
        TF = (DTCM.T / DTCM.T.max()).T
    elif tf_method == 'log':
        TF = (np.log2(DTCM.T + 1)).T
    elif tf_method == 'raw':
        TF = DTCM
    elif tf_method == 'bool':
        TF = DTCM.astype('bool').astype('int')
    else:
        raise ValueError(f"TF method {tf_method} not found.")

    DF = DTCM.count() # Assumes NULLs
    N_docs = len(DTCM)

    if df_method == 'standard':
        IDF = np.log10(N_docs/DF) # This what the students were asked
    elif df_method == 'textbook':
        IDF = np.log10(N_docs/(DF + 1))
    elif df_method == 'sklearn':
        IDF = np.log10(N_docs/DF) + 1
    elif df_method == 'sklearn_smooth':
        IDF = np.log10((N_docs + 1)/(DF + 1)) + 1
    else:
        raise ValueError(f"DF method {df_method} not found.")

    TFIDF = TF * IDF

    DFIDF = DF * IDF

    TFIDF = TFIDF.fillna(0)

    return TFIDF, DFIDF
```

In [116]:
```
OHCO = ['doc_id', 'para_num', 'sent_num', 'token_num']
```

In [117]:
```
SENTS = OHCO[:3]
PARAS = OHCO[:2]
DOCS = OHCO[:1]
```

In [118]:
```
BOW_books = create_bow(CORPUS, bag=DOCS)
```

In [119]:
```
TFIDF_docs, DFIDF_docs = get_tfidf(BOW_books, tf_method='max', df_met
```

In [365]:
```
TFIDF_docs
```

Out[365]:

term_str		aa	aaa	aaaa	aaaaaadaaag	aaanaaaag	aaax	aad	
doc_id									
1	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
2	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
3	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
4	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
5	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
6	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
7	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.05851	0.000000	0.
8	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
9	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
10	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
11	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
12	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
13	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.
14	0.0	0.013122	0.009763	0.026243	0.013122	0.013122	0.00000	0.013122	0.
15	0.0	0.000000	0.059915	0.000000	0.000000	0.000000	0.00000	0.000000	0.

15 rows Ч 4499 columns

In [364]:
```
TFIDF_docs.mean().sort_values(ascending=False)\
    .head(20).to_frame('mean_tfidf').join(VOCAB.max_pos).style.backgr
```

Out[364]:

term_str	mean_tfidf	max_pos
constants	0.040556	NNS
elastic	0.032070	JJ
elasticity	0.037104	NN
magnetic	0.041073	JJ
milling	0.062640	NN
misfit	0.030958	NN
mn	0.028162	NNP
mno	0.049797	NNP
modulus	0.030140	NN
nbmotaw	0.028649	NNP
nico	0.027717	NNP
nicofemncr	0.038595	NNP
nm	0.039823	NN

term_str	mean_tfidf	max_pos
rhea	0.044201	NNP
sigma	0.027422	NN
sss	0.028081	NNP
superconducting	0.028870	VBG

In [121]:
```
BOW_paras = create_bow(CORPUS, bag=PARAS)
```

In [122]:
```
TFIDF_paras_max, DFIDF_paras_max = get_tfidf(BOW_paras, tf_method='ma


TFIDF_paras_max.mean().sort_values(ascending=False)\
    .head(20).to_frame('mean_tfidf').join(VOCAB.max_pos)
```

Out[122]:

term_str	mean_tfidf	max_pos
	0.066817	NaN
alloy	0.052639	NN
alloys	0.062778	NNS
bcc	0.052014	NNP
c	0.044079	NN
cr	0.039821	NNP
entropy	0.041968	NN
fcc	0.039441	NN
hardness	0.049637	NN
heas	0.057680	NNP
high	0.050126	JJ
lattice	0.057338	NN
mechanical	0.040899	JJ
phase	0.054444	NN
properties	0.043947	NNS
rhea	0.039430	NNP
solid	0.041903	JJ
strength	0.042814	NN
temperature	0.045015	NN
v	0.040659	NNP

In [123]:
```
VOCAB['dfidf'] = DFIDF_docs
VOCAB['mean_tfidf'] = TFIDF_docs.mean()
```

In [124]:
```
VOCAB
```

Out[124]:

	n	n_chars	p	i	max_pos	dfidf	mean_tfidf

term_str							
NaN	2580	0	0.081028	3.625443	CD	NaN	NaN
aa	1	2	0.000031	14.958598	JJ	1.176091	0.000234
aaa	3	3	0.000094	13.373636	NNP	1.750123	0.003166
aaaa	2	4	0.000063	13.958598	NNP	1.176091	0.000468
aaaaaadaaag	1	11	0.000031	14.958598	NNP	1.176091	0.000234
...	...	...	...	...	...	...	...
zrrez	1	5	0.000031	14.958598	NNP	1.176091	0.000233
zs	2	2	0.000063	13.958598	NNP	1.750123	0.000570
ztnb	1	4	0.000031	14.958598	NNP	1.176091	0.000233
zwick	1	5	0.000031	14.958598	NNP	1.176091	0.000234
zz	5	2	0.000157	12.636670	NNP	1.176091	0.001170

In [125]:
```
TFIDF_docs[VOCAB.sort_values('n', ascending=False).head(200).sample(1
```

Out[125]:

term_str	phases	composition	elemental	one	ti	nm	shear	
doc_id								
10	0.000000	0.000000	0.001400	0.000956	0.002041	0.000000	0.000000	0.00049
3	0.016494	0.003805	0.000000	0.002537	0.008121	0.000000	0.000000	0.03560
2	0.001772	0.004089	0.003592	0.005724	0.000000	0.025112	0.000000	0.00127
12	0.002405	0.001850	0.003250	0.001295	0.000000	0.000000	0.000000	0.00778
11	0.000000	0.004740	0.002314	0.001053	0.013489	0.000000	0.012130	0.00164
13	0.000916	0.000423	0.000000	0.000423	0.000000	0.000000	0.009737	0.00000
1	0.000000	0.003271	0.000000	0.006542	0.000000	0.018834	0.000000	0.00255
9	0.000000	0.000000	0.000000	0.001053	0.000000	0.000000	0.000000	0.01642
14	0.002010	0.002597	0.000815	0.000557	0.021382	0.000000	0.012818	0.00028
15	0.006908	0.017529	0.000000	0.006374	0.000000	0.048936	0.000000	0.00000

Create DOC

In [126]:
```
LTR
```

Out[126]:

	title	authors	publisher
doc_id			
1	Quantitative determination of the lattice cons...	Zhijun Wang, Qingfeng Wu, Wenquan Zhou, Feng H...	Scripta Materialia 162 (2019) 468-471
2	High entropy multicomponent WMoNbZrV alloy pro...	Dariusz Oleszak, Anna Antolak-Dudka, Tadeusz K...	Materials Letters 232 (2018) 160-162
3	Mapping the magnetic transition temperatures f...	Shuo Huang, Erik Holmstrцm, Olle Eriksson, Lev...	Intermetallics 95 (2018) 80-84

doc_id	title	authors	publisher
4	Mechano-chemical synthesis, thermal stability ...	Vikas Shivam, Joysurya Basu, Yagnesh Shadangi,...	Journal of Alloys and Compounds 757 (2018) 87-97
5	First-principles-based prediction of yield str...	Binglun Yin, William A. Curtin	npj Computational Materials (2019) 5:14
6	Structure and properties of equiatomic CoCrFeN...	A.S. Rogachev, S.G. Vadchenko, N.A. Kochetov, ...	Journal of Alloys and Compounds 805 (2019) 123...
7	Assessing elastic property and solid-solution ...	Zhi-biao Yang, Jian Sun	Journal of Materials Research volume 33, pages...
8	Fast production of high entropy alloys (CoCrFe...	Azmi Erdogan, Tuba Yener, Sakin Zeytin	Vacuum 155 (2018) 64-72
9	Chemical complexity induced local structural d...	Fuxiang Zhang, Yang Tong, Ke Jin, Hongbin Bei,...	Materials Research Letters, 6:8, 450-455
10	Microstructures and properties of Al0.3CoCrFeN...	Sze-Kwan Wong, Tao-Tsung Shun, Chieh-Hsiang Ch...	Materials Chemistry and Physics 210 (2018) 146...
11	Vanadium is an optimal element for strengtheni...	Binglun Yin, Francesco Maresca, W.A. Curtin	Acta Materialia 188 (2020) 486-491
12	High-entropy alloy superconductors on an α-Mn ...	Karoline Stolze, F. Alex Cevallos, Tai Kong, R...	Journal of Materials Chemistry C, 2018, 6, 10441
13	First-principle calculation investigation of N...	Y.L. Hu, L.H. Bai, Y.G. Tong, D.Y. Deng, X.B. ...	Journal of Alloys and Compounds 827 (2020) 153963

In [127]:
```
lib_cols = "title authors publisher".split()
DOC = pd.DataFrame(index=TFIDF_docs.index)
DOC = DOC.join(LIB[lib_cols], on='doc_id')
```

In [128]:
```
DOC
```

Out[128]:

doc_id	title	authors	publisher
1	Quantitative determination of the lattice cons...	Zhijun Wang, Qingfeng Wu, Wenquan Zhou, Feng H...	Scripta Materialia 162 (2019) 468-471
2	High entropy multicomponent WMoNbZrV alloy pro...	Dariusz Oleszak, Anna Antolak-Dudka, Tadeusz K...	Materials Letters 232 (2018) 160-162
3	Mapping the magnetic transition temperatures f...	Shuo Huang, Erik Holmstrцm, Olle Eriksson, Lev...	Intermetallics 95 (2018) 80-84
4	Mechano-chemical synthesis, thermal stability ...	Vikas Shivam, Joysurya Basu, Yagnesh Shadangi,...	Journal of Alloys and Compounds 757 (2018) 87-97
5	First-principles-based prediction of yield str...	Binglun Yin, William A. Curtin	npj Computational Materials (2019) 5:14
6	Structure and properties of equiatomic CoCrFeN...	A.S. Rogachev, S.G. Vadchenko, N.A. Kochetov, ...	Journal of Alloys and Compounds 805 (2019) 123...

doc_id	title	authors	publisher
7	Assessing elastic property and solid-solution ...	Zhi-biao Yang, Jian Sun	Journal of Materials Research volume 33, pages...
8	Fast production of high entropy alloys (CoCrFe...	Azmi Erdogan, Tuba Yener, Sakin Zeytin	Vacuum 155 (2018) 64-72
9	Chemical complexity induced local structural d...	Fuxiang Zhang, Yang Tong, Ke Jin, Hongbin Bei,...	Materials Research Letters, 6:8, 450-455
10	Microstructures and properties of Al0.3CoCrFeN...	Sze-Kwan Wong, Tao-Tsung Shun, Chieh-Hsiang Ch...	Materials Chemistry and Physics 210 (2018) 146...
11	Vanadium is an optimal element for strengtheni...	Binglun Yin, Francesco Maresca, W.A. Curtin	Acta Materialia 188 (2020) 486-491
12	High-entropy alloy superconductors on an α-Mn ...	Karoline Stolze, F. Alex Cevallos, Tai Kong, R...	Journal of Materials Chemistry C, 2018, 6, 10441
13	First-principle calculation investigation of N...	Y.L. Hu, L.H. Bai, Y.G. Tong, D.Y. Deng, X.B. ...	Journal of Alloys and Compounds 827 (2020) 153963

In [129]:
```
DOC['mean_tfidf'] = TFIDF_docs.mean(1)
DOC['n_tokens'] = BOW_books.groupby(['doc_id']).n.sum()
```

In [130]:
```
DOC
```

Out[130]:

doc_id	title	authors	publisher	mean_tfidf	n_tokens
1	Quantitative determination of the lattice cons...	Zhijun Wang, Qingfeng Wu, Wenquan Zhou, Feng H...	Scripta Materialia 162 (2019) 468-471	0.001416	1385
2	High entropy multicomponent WMoNbZrV alloy pro...	Dariusz Oleszak, Anna Antolak-Dudka, Tadeusz K...	Materials Letters 232 (2018) 160-162	0.000871	903
3	Mapping the magnetic transition temperatures f...	Shuo Huang, Erik Holmstrцm, Olle Eriksson, Lev...	Intermetallics 95 (2018) 80-84	0.002478	1258
4	Mechano-chemical synthesis, thermal stability ...	Vikas Shivam, Joysurya Basu, Yagnesh Shadangi,...	Journal of Alloys and Compounds 757 (2018) 87-97	0.001609	3320
5	First-principles-based prediction of yield str...	Binglun Yin, William A. Curtin	npj Computational Materials (2019) 5:14	0.001437	3187
6	Structure and properties of equiatomic CoCrFeN...	A.S. Rogachev, S.G. Vadchenko, N.A. Kochetov, ...	Journal of Alloys and Compounds 805 (2019) 123...	0.000842	2717
7	Assessing elastic property and solid-solution ...	Zhi-biao Yang, Jian Sun	Journal of Materials Research volume 33, pages...	0.000781	2569
8	Fast production of high entropy alloys (CoCrFe...	Azmi Erdogan, Tuba Yener, Sakin Zeytin	Vacuum 155 (2018) 64-72	0.001318	1984

doc_id	title	authors	publisher	mean_tfidf	n_tokens
9	Chemical complexity induced local structural d...	Fuxiang Zhang, Yang Tong, Ke Jin, Hongbin Bei,...	Materials Research Letters, 6:8, 450-455	0.001165	1475
10	Microstructures and properties of Al0.3CoCrFeN...	Sze-Kwan Wong, Tao-Tsung Shun, Chieh-Hsiang Ch...	Materials Chemistry and Physics 210 (2018) 146...	0.000522	1226
11	Vanadium is an optimal element for strengtheni...	Binglun Yin, Francesco Maresca, W.A. Curtin	Acta Materialia 188 (2020) 486-491	0.001527	1999
12	High-entropy alloy superconductors on an α-Mn ...	Karoline Stolze, F. Alex Cevallos, Tai Kong, R...	Journal of Materials Chemistry C, 2018, 6, 10441	0.000942	2932
13	First-principle calculation investigation of N...	Y.L. Hu, L.H. Bai, Y.G. Tong, D.Y. Deng, X.B. ...	Journal of Alloys and Compounds 827 (2020) 153963	0.001339	1977
14	Contribution of Lattice Distortion to Solid So...	H. Chen, A. Kauffmann, S. Laube, I.-C. Choi, R...	Metallurgical and Materials Transactions A vol...	0.001260	3900

Compute PCA

In [131]:
```
norm_docs = True # L2 norming
center_by_mean = False
center_by_variance = False # Not supposed to ... Exaggerates signific
n_terms = 1000 # Number of significant words; feature space
k = 6 # Number of components
from scipy.linalg import norm
import plotly.express as px
```

In [132]:
```
if norm_docs:
    print("L2 norming")
    TFIDF_docs = TFIDF_docs.apply(lambda x: x / norm(x), 1).fillna(0)
```

```
L2 norming
```

In [133]:
```
if center_by_mean:
    print("Centering by mean")
    TFIDF_docs = TFIDF_docs - TFIDF_docs.mean()
```

In [134]:
```
if center_by_variance:
    print("Centering by variance")
    TFIDF_docs = TFIDF_docs / TFIDF_docs.std()
```

In [135]:
```
COV = TFIDF_docs.cov()
```

In [136]:
```
COV
```

Out[136]:

term_str	aa	aaa	aaaa	aaaaaadaaag	aaanaaaag	a
term_str						
0.0	0.000000e+00	0.000000	0.000000	0.000000e+00	0.000000e+00	0.000
aa 0.0	1.147847e-05	0.000005	0.000023	1.147847e-05	1.147847e-05	-0.000
aaa 0.0	4.796720e-06	0.000240	0.000010	4.796720e-06	4.796720e-06	-0.000

term_str		aa	aaa	aaaa	aaaaaadaaag	aaanaaaag	a
term_str							
aaaa	0.0	2.295693e-05	0.000010	0.000046	2.295693e-05	2.295693e-05	-0.000
aaaaaadaaag	0.0	1.147847e-05	0.000005	0.000023	1.147847e-05	1.147847e-05	-0.000
...	...	...	...	...	...	...	
zrrez	0.0	-8.199548e-07	-0.000004	-0.000002	-8.199548e-07	-8.199548e-07	-0.000
zs	0.0	-1.180895e-06	-0.000006	-0.000002	-1.180895e-06	-1.180895e-06	-0.000
ztnb	0.0	-8.199548e-07	-0.000004	-0.000002	-8.199548e-07	-8.199548e-07	-0.000
zwick	0.0	1.147847e-05	0.000005	0.000023	1.147847e-05	1.147847e-05	-0.000

In [137]:
```
COV.stack().sort_values().loc['lattice']
```

Out[137]:
```
term_str
although       0.0
aluminium      0.0
aluminum       0.0
aly            0.0
ambient        0.0
               ...
embrittling    0.0
eaa            0.0
due            0.0
duplex         0.0
ductility      0.0
Length: 4499, dtype: float64
```

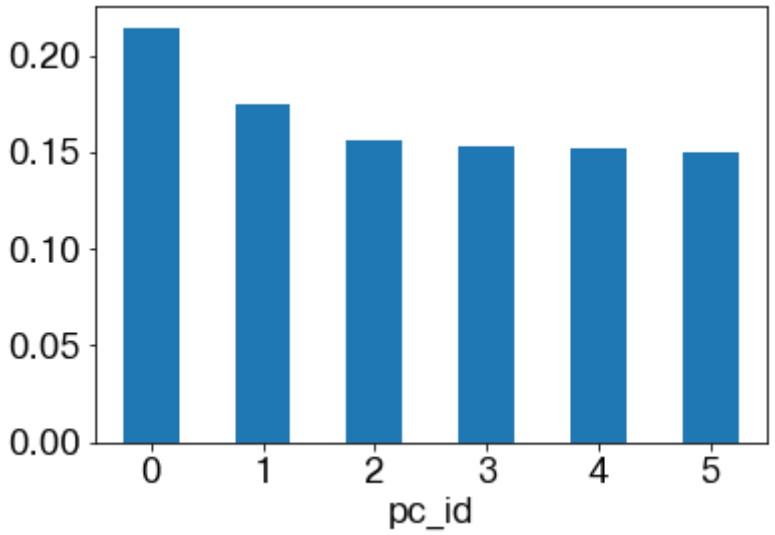
In [138]:
```
from scipy.linalg import eigh
```

In [139]:
```
w, v = eigh(COV, subset_by_index=(len(COV)-k, len(COV)-1))
```

In [140]:
```
pc_ids = list(reversed([i for i in range(k)]))
```

In [141]:
```
C = pd.DataFrame(w, columns=['eig_val'])
C.index = pc_ids
C = C.sort_index()
C.index.name = 'pc_id'
C['exp_var'] = C.eig_val / C.eig_val.sum()
```

In [142]:
```
C_exp_var_plot_bar(rot=0);
```



In [143]:
```
L = pd.DataFrame(v, index=COV.index)
L.columns = pc_ids
L = L.T.sort_index().T
L.columns.name = 'pc_id'
```

In [144]:
```
L.sample(20).style.background_gradient(cmap='GnBu', high=.5)
```

Out[144]:

pc_id	0	1	2	3	4	5
term_str						
lnimn	0.003767	0.000161	0.001105	-0.001178	-0.003291	-0.000309
representation	0.000555	-0.003423	-0.002884	-0.002411	0.004222	-0.005271
mj	-0.002361	-0.004654	-0.018522	-0.005967	0.008593	0.001032
recorded	0.020382	0.017555	-0.002650	0.005222	0.000766	-0.003146
clean	0.002384	-0.003076	0.006978	-0.006876	0.002910	0.006580
power	-0.003339	0.001309	-0.020055	-0.005862	0.009003	0.000136
zeroresistance	0.001111	-0.006846	-0.005768	-0.004823	0.008443	-0.010543
corner	0.000171	-0.003148	-0.000148	-0.004547	0.003303	-0.001240
wre	0.000555	-0.003423	-0.002884	-0.002411	0.004222	-0.005271
nialo	0.002384	-0.003076	0.006978	-0.006876	0.002910	0.006580
economic	-0.000721	-0.002199	0.002988	0.009128	0.004627	-0.001722
equiatomic	0.022433	0.014554	-0.033096	-0.013569	0.006981	0.004881
investigating	-0.000650	-0.001617	0.005370	-0.007401	0.000434	0.007210
adjacent	-0.007996	0.006217	-0.002547	0.001112	-0.006374	-0.000512
fee	-0.000325	-0.000808	0.002685	-0.003700	0.000217	0.003605
macroscopically	-0.000325	-0.000808	0.002685	-0.003700	0.000217	0.003605
strain	0.020922	0.023648	0.002220	0.002082	-0.001049	-0.000058
computation	-0.011416	0.011997	0.004365	0.000496	-0.000199	-0.000978
multiprincipal	-0.008034	0.001871	-0.006119	0.008647	-0.009157	-0.001577
rapidly	-0.003834	0.001916	-0.000522	-0.001610	0.003067	-0.004286

In [145]:
```
DCM = TFIDF_docs.dot(L)
```

In [146]:
```
top_n = 6 # Number of top words for each pole
for i in range(k):
    for j, pole in enumerate(['neg','pos']):
        top_terms = ' '.join(L.sort_values(i, ascending=bool(j)).head
        C.loc[i, pole] = top_terms
```

In [147]:
```
try:
    DOC = DOC.join(DCM)
except:
    pass
```

In [148]:
```
DOC['authors'] = DOC['authors'].apply(lambda x: x.split(' ')[0]+' '+
```

In [149]:
```
DOC["title"] = DOC[["title", "publisher"]].apply("\n".join, axis=1)
DOC.title
```

Out[149]:

```
doc_id
1     Quantitative determination of the lattice cons...
2     High entropy multicomponent WMoNbZrV alloy pro...
3     Mapping the magnetic transition temperatures f...
4     Mechano-chemical synthesis, thermal stability ...
5     First-principles-based prediction of yield str...
6     Structure and properties of equiatomic CoCrFeN...
7     Assessing elastic property and solid-solution ...
8     Fast production of high entropy alloys (CoCrFe...
9     Chemical complexity induced local structural d...
10    Microstructures and properties of Al0.3CoCrFeN...
11    Vanadium is an optimal element for strengtheni
```

In [516]:
```
px.scatter(DOC, 0, 1,
           color='authors',
           hover_name='title',
           height=1000, marginal_x='box', marginal_y='box')
```

In [151]:
```
try:
    VOCAB2 = VOCAB.join(L, how='right').reset_index()
except:
    pass
```

VOCAB['dfidf'] = DFIDF_docs VOCAB['mean_tfidf'] = TFIDF_docs.mean()

In [155]:
```
VOCAB2.dropna(inplace=True)
```

In [156]:
```
VOCAB2
```

Out[156]:

	term_str	n	n_chars	p	i	max_pos	dfidf	mean_tfidf	
1	aa	1.0	2.0	0.000031	14.958598	JJ	1.176091	0.000234	-0.00
2	aaa	3.0	3.0	0.000094	13.373636	NNP	1.750123	0.003166	-0.00
3	aaaa	2.0	4.0	0.000063	13.958598	NNP	1.176091	0.000468	-0.00
4	aaaaaadaaag	1.0	11.0	0.000031	14.958598	NNP	1.176091	0.000234	-0.00
5	aaanaaaag	1.0	9.0	0.000031	14.958598	NNP	1.176091	0.000234	-0.00
...	...	...	...	...	...	...	...	...	
4494	zrrez	1.0	5.0	0.000031	14.958598	NNP	1.176091	0.000233	0.00
4495	zs	2.0	2.0	0.000063	13.958598	NNP	1.750123	0.000570	-0.00
4496	ztnb	1.0	4.0	0.000031	14.958598	NNP	1.176091	0.000233	0.00
4497	zwick	1.0	5.0	0.000031	14.958598	NNP	1.176091	0.000234	-0.00
4498	zz	5.0	2.0	0.000157	12.636670	NNP	1.176091	0.001170	-0.00

4496 rows Ч 14 columns

In [517]:
```
px.scatter(VOCAB2, 0, 1,
           size='mean_tfidf',color='dfidf',
           hover_name='term_str',
           hover_data=['max_pos'],
           marginal_x='box', marginal_y='box',
           height=1000, width=1000)
```

In [158]:
```
px.scatter(VOCAB2, 2, 3,
           size='mean_tfidf',color='dfidf',
           hover_name='term_str', hover_data=['max_pos'],
           marginal_x='box', marginal_y='box',
           height=1000, width=1000)
```

In [159]:
```
C
```

Out[159]:

pc_id	eig_val	exp_var	neg	pos
0	0.099633	0.214659	milling crystallite powders milled nm upto	volumes constants misfit elastic theory apparent
1	0.081015	0.174546	milling volumes misfit apparent crystallite co...	mno nicofemncr rhea superconducting zrnb elast...
2	0.072310	0.155792	mno rhea sigma gray elasticity nbmotaw	magnetic tc nicofemncr sss nico ternary
3	0.071054	0.153086	rhea nbmotaw nbmotawv nicofemncr band vca	sigma nbmocrtial gray mno cocrfenial ti
4	0.070504	0.151901	rhea nbmotaw superconducting nbmotawv magnetic...	mno nicofemncr exafs debye pitting nd
5	0.069629	0.150016	nicofemncr sigma exafs elasticity debye nbmocr...	mno superconducting pitting zrnb sus hftawpt

In [160]:
```
BOW_books
```

Out[160]:

doc_id	term_str	n
		76
	ab	5
1	abinitial	1
	ac	3
	accordingly	2
...	...	...
	works	1
	wow	1
15	x	1
	xpixture	1
	yield	2

10518 rows Ч 1 columns

TopicModel

In [161]:
```
from topicmodel import TopicModel
```

In [217]:
```
n_topics = 10
n_terms = 4000
```

In [218]:
```
BOW = BOW_books
```

In [219]:
```
tm = TopicModel(BOW)
```

In [220]:
```
tm.n_topics = n_topics
```

```
tm.n_terms = n_terms
```

In [221]:
```
tm.create_X()
```

In [222]:
```
tm.get_model()
tm.describe_topics()
tm.get_model_stats()
```

In [223]:
```
tm.THETA.sum().idxmax()
```

Out[223]: 2

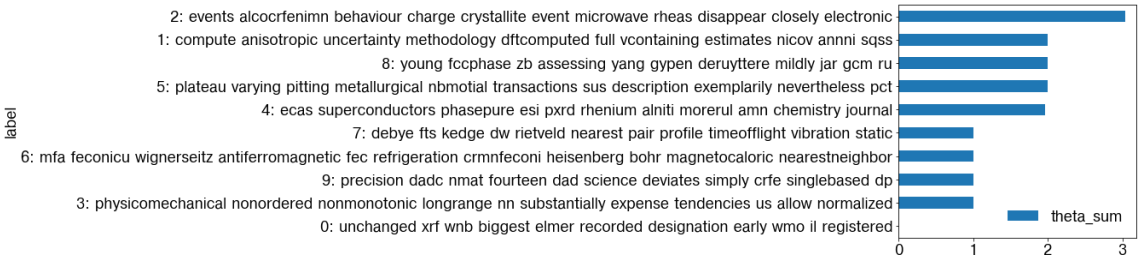
In [224]:
```
tm.PHI.sum(1).idxmax()
```

Out[224]: 2

In [225]:
```
import matplotlib.pyplot as plt
```

In [226]:
```
plt.figure(figsize=(40,10))
tm.plot_topics()
```

```
<Figure size 2880x720 with 0 Axes>
```



In [172]:
```
tm.PHI
```

Out[172]:

term_str	hf	match	needed	depends	derived	note	sqss	spectrum	spectroscopy	spe
topic_id										
0	0.1	0.1	0.100000	0.1	0.100000	0.1	0.1	0.1	0.100000	0.
1	3.1	1.1	3.100000	3.1	1.100000	2.1	6.1	0.1	0.100000	1.
2	0.1	3.1	0.100002	0.1	0.100002	3.1	0.1	5.1	1.099999	1.
3	0.1	0.1	0.100000	1.1	0.100000	0.1	0.1	0.1	2.100001	0.
4	3.1	2.1	1.099998	0.1	1.099998	0.1	0.1	0.1	0.100000	2.
5	0.1	0.1	1.100000	0.1	1.100000	0.1	0.1	0.1	3.100000	1.
6	0.1	0.1	0.100000	0.1	0.100000	0.1	0.1	0.1	0.100000	0.
7	0.1	0.1	0.100000	0.1	2.100001	0.1	0.1	1.1	0.100000	1.
8	0.1	0.1	1.100000	2.1	1.100000	1.1	0.1	0.1	0.100000	0.
9	0.1	0.1	0.100000	0.1	0.100000	0.1	0.1	0.1	0.100000	0.

10 rows Ч 4000 columns

In [173]:
```
tm.THETA
```

Out[173]:

topic_id	0	1	2	3	4	5	6	7
doc_id								

topic_id	0	1	2	3	4	5	6	7	
doc_id									
1	0.000256	0.000256	0.000256	0.000256	0.000256	0.000256	0.000256	0.000256	0
2	0.000364	0.000364	0.996727	0.000364	0.000364	0.000364	0.000364	0.000364	0
3	0.000202	0.000202	0.000202	0.000202	0.000202	0.000202	0.998178	0.000202	0
4	0.000093	0.000093	0.999161	0.000093	0.000093	0.000093	0.000093	0.000093	0
5	0.000090	0.999192	0.000090	0.000090	0.000090	0.000090	0.000090	0.000090	0
6	0.000108	0.000108	0.000108	0.000108	0.000108	0.000108	0.000108	0.000108	0
7	0.000197	0.000197	0.000197	0.000197	0.000197	0.000197	0.000197	0.000197	0
8	0.000182	0.000182	0.038596	0.000182	0.959947	0.000182	0.000182	0.000182	0
9	0.000195	0.000195	0.000195	0.000195	0.000195	0.000195	0.000195	0.998242	0
10	0.000281	0.000281	0.000281	0.000281	0.000281	0.997472	0.000281	0.000281	0
11	0.000165	0.998512	0.000165	0.000165	0.000165	0.000165	0.000165	0.000165	0
12	0.000093	0.000093	0.000093	0.000093	0.999160	0.000093	0.000093	0.000093	0

In [174]:
```
tm.TOPIC
```

Out[174]:

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	la
0	400.000000	0.002827	11.97	unchanged xrf wnb biggest elmer recorded desig...	mesh consists widely confirmed concept microst...	0: unchanged wnb biggest eln recorded d
1	2117.000003	2.000277	10.49	compute anisotropic uncertainty methodology df...	compute anisotropic uncertainty across predict...	1: comp anisotro uncertai methodolog
2	2344.132801	3.034942	10.53	events alcocrfenimn behaviour charge crystalli...	events electronic alcocrfenimn behaviour charg...	2: eve alcocrfeni behaviour chai cryst
3	676.000023	0.999217	10.99	physicomechanical nonordered nonmonotonic long...	normalized physicomechanical demonstrates us i...	physicomechani nonorder nonmonotonic
4	1996.867094	1.961658	10.51	ecas superconductors phasepure esi pxrd rheniu...	ecas superconductors phasepure areas esi magne...	4: ec superconduct phasepure pxrd rh
5	2266.999982	1.999357	10.67	plateau varying pitting metallurgical nbmotial...	plateau varying pitting nbmotial metallurgical...	5: plateau vary pitt metallurgi nbmc
6	893.000028	1.000803	10.77	mfa feconicu wignerseitz antiferromagnetic fec...	curie exchange ferromagnetic feconicu wignerse...	6: mfa fecon wigners antiferromagne
7	911.000043	1.000874	10.53	debye fts kedge dw rietveld nearest pair profi...	debye fts distance neutron vibration static fi...	7: debye kedge dw rietv nearest pair p

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	la
8	1834.999993	1.999780	10.61	young fccphase zb assessing yang	young fccphase assessing yang zb	8: young fccpha zb assessing ya

In [175]:
```
from gensim.models import word2vec
from sklearn.manifold import TSNE
```

In [176]:
```
from w2v import W2V
```

In [177]:
```
HEAs = W2V(CORPUS, OHCO[:-1], OHCO[:2])
HEAs.w2v_args['min_count'] = 50
HEAs.w2v_args['vector_size'] = 256
HEAs.tsne_args['perplexity'] = 20
HEAs.tsne_args['random_state'] = 111
HEAs.generate_model();
```

```
W2V Bag: sent_num
DOC Bag: para_num
Extracting vocabulary
Gathering sentences
Learning word vectors
Computing tSNE coordinates
Done ✓
```

In [178]:
```
HEAs.TSNE
```

Out[178]:

term_str	x	y	n	dfidf	pos_group
	-454.805542	-47.095867	2580	79.736942	CD
addition	-185.773575	26.017103	116	142.212808	NN
al	166.479248	554.300354	98	133.206067	NN
alloy	14.078995	-13.189383	376	149.919653	NN
alloying	-170.668045	126.097061	76	134.311243	VB
...	...	...	...	...	...
well	-124.588707	-428.994507	52	119.086256	RB
work	367.958160	231.045059	58	124.660879	NN
x	22.495687	633.815918	122	124.660879	JJ
xrd	247.893005	39.435860	78	124.660879	NN
yield	90.229935	27.087635	66	109.648414	NN

105 rows Ч 5 columns

In [366]:
```
px.scatter(HEAs.TSNE.reset_index(), 'x', 'y',
           color='pos_group', text='term_str',
           hover_name='dfidf',
           size=np.log2(HEAs.TSNE.n),
           height=1000, width=1200, marginal_x='box', marginal_y='box'
```

In [181]:
```
HEAs_TSNE['n2'] = np.log2(HEAs_TSNE.n)
```

In [228]:
```
HEAs.plot_tsne(n=2000, method='dfidf')
```

In [227]:
```
HEAs.plot_tsne?
```

In [183]:
```
HEAs.complete_analogy('structure', 'properties', 'fcc', 10)
```

Out[183]:

	term	sim
0	high	0.998927
1	alloys	0.998924
2	energy	0.998903
3	heas	0.998899
4	modulus	0.998888
5	lattice	0.998884
6	v	0.998875
7	calculated	0.998851
8	effect	0.998847
9	mn	0.998817

In [184]:
```
HEAs.complete_analogy('structure', 'properties', 'bcc', 10)
```

Out[184]:

	term	sim
0	mn	0.998920
1	heas	0.998912
2	elements	0.998912
3	alloys	0.998895
4	calculated	0.998888
5	phases	0.998888
6	addition	0.998886
7	crystal	0.998876
8	ni	0.998864
9	modulus	0.998863

In [186]:
```
HEAs.get_most_similar('bcc')
```

Out[186]:

	term	sim
0	phases	0.999451
1	structure	0.999436
2	alloys	0.999430
3	elements	0.999426
4	calculated	0.999408
5	alloy	0.999408
6	addition	0.999404
7	solid	0.999392
8	lattice	0.999384

term sim

Sentiment Analysis

In [303]:
```
emo_cols = "optimal determination strength evolution calculation comm
```

In [231]:
```
from IPython.display import display, HTML
```

In [342]:
```
SALEX = pd.read_csv('salex_combo.csv').drop_duplicates(subset=['term_
if len(set([idx.lower() for idx in SALEX.index])) == len(SALEX.index)
    SALEX.index = [idx.lower() for idx in SALEX.index]
assert SALEX.index.is_unique
```

In [346]:
```
SALEX=pd.read_csv('salex_syuzhet.csv').set_index('term_str')
```

In [347]:
```
SALEX.columns = [col.replace('syu_', '') for col in SALEX.columns]
```

In [348]:
```
SALEX
```

Out[348]:

term_str	sentiment
abandon	-0.75
abandoned	-0.50
abandoner	-0.25
abandonment	-0.25
abandons	-1.00
...	...
zest	0.50
zombie	-0.25
zombies	-0.25
false	-0.60
true	0.50

10747 rows Ч 1 columns

In [349]:
```
try:
    VOCAB3 = VOCAB.join(SALEX)
except ValueError:
    pass
```

In [351]:
```
VOCAB3.dropna(inplace=True)
```

In [352]:
```
VOCAB3
```

Out[352]:

term_str	n	n_chars	p	i	max_pos	dfidf	mean_tfidf	sentiment

term_str	n	n_chars	p	i	max_pos	dfidf	mean_tfidf	sentiment
aberration	1	10	0.000031	14.958598	NN	1.176091	0.000259	-0.80
abilities	2	9	0.000063	13.958598	NNS	1.750123	0.000794	0.60
ability	5	7	0.000157	12.636670	NN	2.385606	0.000668	0.50
abnormal	1	8	0.000031	14.958598	JJ	1.176091	0.000234	-0.50
abrasive	1	8	0.000031	14.958598	JJ	1.176091	0.000234	-0.50
...	...	...	...	...	...	...	...	...
worth	3	5	0.000094	13.373636	JJ	1.750123	0.000823	0.75
worthwhile	1	10	0.000031	14.958598	JJ	1.176091	0.000234	1.00
worthy	1	6	0.000031	14.958598	JJ	1.176091	0.000424	0.75
wow	1	3	0.000031	14.958598	NN	1.176091	0.002010	0.50

In [ ]:
```
DOC_new = LIB.join(LIB.doc_id).groupby('city_id').sum().T
CITY_TERMS.index.name = 'term_str'
```

In [319]:
```
CORPUS
```

Out[319]:

doc_id	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str
			1	(variation, NN)	NN	variation	variation
			4	(lattice, NN)	NN	lattice	lattice
1	0	0	5	(constant, NN)	NN	constant	constant
			7	(alloying, VBG)	VBG	alloying	alloying
			8	(elements, NNS)	NNS	elements	elements
...	...	...	...	...	...	...	...
			22	(based, VBN)	VBN	based	based
			25	(role, NN)	NN	role	role
15	20	0	28	(refractory, JJ)	JJ	refractory	refractory
			29	(metal, NN)	NN	metal	metal
			32	(composition, NN)	NN	composition	composition

31841 rows Ч 4 columns

In [359]:
```
COMBO = CORPUS.join(LIB).join(SALEX, on='term_str').join(BOW_books, c
COMBO = COMBO.drop(['n'], axis=1)
COMBO = COMBO.sort_index()
```

In [360]:
```
BOW_books
```

Out[360]:

doc_id	term_str	n
1		76
	ab	5

		n
doc_id	term_str	
	abinitial	1
	ac	3
	accordingly	2
...	...	...
	works	1
	wow	1
15	x	1
	xpixture	1

In [361]:
```
COMBO
```

Out[361]:

				pos_tuple	pos	token_str	term_str	
doc_id	para_num	sent_num	token_num					
			1	(variation, NN)	NN	variation	variation	Qua determ of the
			4	(lattice, NN)	NN	lattice	lattice	Qua determ of the
1	0	0	5	(constant, NN)	NN	constant	constant	Qua determ of the
			7	(alloying, VBG)	VBG	alloying	alloying	Qua determ of the
			8	(elements, NNS)	NNS	elements	elements	Qua determ of the
...	...	...	...	...	...	...	...	

doc_id	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str	
			22	(based, VBN)	VBN	based	based	Para Fo
			25	(role, NN)	NN	role	role	Para Fo
15	20	0	28	(refractory, JJ)	JJ	refractory	refractory	Para Fo
			29	(metal, NN)	NN	metal	metal	Para Fo

In [362]:
```
DOCS = COMBO.groupby(OHCO[:1])[emo_cols].mean()
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-362-9239d18715a3> in <module>
----> 1 DOCS = COMBO.groupby(OHCO[:1])[emo_cols].mean()

~/anaconda3/lib/python3.8/site-packages/pandas/core/groupby/generic.py in __getitem__(self, key)
   1540                 stacklevel=2,
   1541             )
-> 1542         return super().__getitem__(key)
   1543
   1544     def _gotitem(self, key, ndim: int, subset=None):

~/anaconda3/lib/python3.8/site-packages/pandas/core/base.py in __getitem__(self, key)
    266                     set(key).difference(self.obj.columns)  # type: ignore[attr-defined]
    267                 )
--> 268                 raise KeyError(f"Columns not found: {str(bad_keys)[1:-1]}")
    269             return self._gotitem(list(key), ndim=2)
    270

KeyError: "Columns not found: 'complexity', 'calculation', 'optimal', 'evolution', 'strength', 'determination'"
```

In [297]:
```
LTR
```

Out[297]:

doc_id	title	authors	publisher
1	Quantitative determination of the lattice cons...	Zhijun Wang, Qingfeng Wu, Wenquan Zhou, Feng H...	Scripta Materialia 162 (2019) 468-471
2	High entropy multicomponent WMoNbZrV alloy pro...	Dariusz Oleszak, Anna Antolak-Dudka, Tadeusz K...	Materials Letters 232 (2018) 160-162
3	Mapping the magnetic transition temperatures f...	Shuo Huang, Erik Holmstrцm, Olle Eriksson, Lev...	Intermetallics 95 (2018) 80-84
4	Mechano-chemical synthesis, thermal stability ...	Vikas Shivam, Joysurya Basu, Yagnesh Shadangi,...	Journal of Alloys and Compounds 757 (2018) 87-97
5	First-principles-based prediction of yield str...	Binglun Yin, William A. Curtin	npj Computational Materials (2019) 5:14
6	Structure and properties of equiatomic CoCrFeN...	A.S. Rogachev, S.G. Vadchenko, N.A. Kochetov, ...	Journal of Alloys and Compounds 805 (2019) 123...
7	Assessing elastic property and solid-solution ...	Zhi-biao Yang, Jian Sun	Journal of Materials Research volume 33, pages...
8	Fast production of high entropy alloys (CoCrFe...	Azmi Erdogan, Tuba Yener, Sakin Zeytin	Vacuum 155 (2018) 64-72
9	Chemical complexity induced local structural d...	Fuxiang Zhang, Yang Tong, Ke Jin, Hongbin Bei,...	Materials Research Letters, 6:8, 450-455
10	Microstructures and properties of Al0.3CoCrFeN...	Sze-Kwan Wong, Tao-Tsung Shun, Chieh-Hsiang Ch...	Materials Chemistry and Physics 210 (2018) 146...
11	Vanadium is an optimal element for strengtheni...	Binglun Yin, Francesco Maresca, W.A. Curtin	Acta Materialia 188 (2020) 486-491
12	High-entropy alloy superconductors on an α-Mn ...	Karoline Stolze, F. Alex Cevallos, Tai Kong, R...	Journal of Materials Chemistry C, 2018, 6, 10441
13	First-principle calculation investigation of N...	Y.L. Hu, L.H. Bai, Y.G. Tong, D.Y. Deng, X.B. ...	Journal of Alloys and Compounds 827 (2020) 153963
14	Contribution of Lattice Distortion to Solid So...	H. Chen, A. Kauffmann, S. Laube, I.-C. Choi, R...	Metallurgical and Materials Transactions A vol...
15	Role of Various Parameters in the Formation of...	V. F. Gorban, N. A. Krapivka, S. A. Firstov, D...	Physics of Metals and Metallography, Volume 11...

In [298]:
```
LTR.title
```

Out[298]:

```
doc_id
1     Quantitative determination of the lattice cons...
2     High entropy multicomponent WMoNbZrV alloy pro...
3     Mapping the magnetic transition temperatures f...
4     Mechano-chemical synthesis, thermal stability ...
5     First-principles-based prediction of yield str...
6     Structure and properties of equiatomic CoCrFeN...
7     Assessing elastic property and solid solution
```

In [300]:
```
LIB.to_csv('LIB_Ankita.csv')
```

In [301]:
```
LIB
```

Out[301]:

doc_id	title	authors	publisher
1	Quantitative determination of the lattice cons...	Zhijun Wang, Qingfeng Wu, Wenquan Zhou, Feng H...	Scripta Materialia 162 (2019) 468-471
2	High entropy multicomponent WMoNbZrV alloy pro...	Dariusz Oleszak, Anna Antolak-Dudka, Tadeusz K...	Materials Letters 232 (2018) 160-162
3	Mapping the magnetic transition temperatures f...	Shuo Huang, Erik Holmstrцm, Olle Eriksson, Lev...	Intermetallics 95 (2018) 80-84
4	Mechano-chemical synthesis, thermal stability ...	Vikas Shivam, Joysurya Basu, Yagnesh Shadangi,...	Journal of Alloys and Compounds 757 (2018) 87-97
5	First-principles-based prediction of yield str...	Binglun Yin, William A. Curtin	npj Computational Materials (2019) 5:14
6	Structure and properties of equiatomic CoCrFeN...	A.S. Rogachev, S.G. Vadchenko, N.A. Kochetov, ...	Journal of Alloys and Compounds 805 (2019) 123...
7	Assessing elastic property and solid-solution ...	Zhi-biao Yang, Jian Sun	Journal of Materials Research volume 33, pages...
8	Fast production of high entropy alloys (CoCrFe...	Azmi Erdogan, Tuba Yener, Sakin Zeytin	Vacuum 155 (2018) 64-72
9	Chemical complexity induced local structural d...	Fuxiang Zhang, Yang Tong, Ke Jin, Hongbin Bei,...	Materials Research Letters, 6:8, 450-455
10	Microstructures and properties of Al0.3CoCrFeN...	Sze-Kwan Wong, Tao-Tsung Shun, Chieh-Hsiang Ch...	Materials Chemistry and Physics 210 (2018) 146...
11	Vanadium is an optimal element for strengtheni...	Binglun Yin, Francesco Maresca, W.A. Curtin	Acta Materialia 188 (2020) 486-491
12	High-entropy alloy superconductors on an α-Mn ...	Karoline Stolze, F. Alex Cevallos, Tai Kong, R...	Journal of Materials Chemistry C, 2018, 6, 10441
13	First-principle calculation investigation of N...	Y.L. Hu, L.H. Bai, Y.G. Tong, D.Y. Deng, X.B. ...	Journal of Alloys and Compounds 827 (2020) 153963
14	Contribution of Lattice Distortion to Solid So...	H. Chen, A. Kauffmann, S. Laube, I.-C. Choi, R...	Metallurgical and Materials Transactions A vol...
15	Role of Various Parameters in the Formation of...	V. F. Gorban, N. A. Krapivka, S. A. Firstov, D...	Physics of Metals and Metallography, Volume 11...

In [312]:
```
VOCAB
```

Out[312]:

term_str	n	n_chars	p	i	max_pos	dfidf	mean_tfidf
NaN	2580	0	0.081028	3.625443	CD	NaN	NaN
aa	1	2	0.000031	14.958598	JJ	1.176091	0.000234
aaa	3	3	0.000094	13.373636	NNP	1.750123	0.003166
aaaa	2	4	0.000063	13.958598	NNP	1.176091	0.000468
aaaaaadaaag	1	11	0.000031	14.958598	NNP	1.176091	0.000234
...	...	...	...	...	...	...	...
zrrez	1	5	0.000031	14.958598	NNP	1.176091	0.000233
zs	2	2	0.000063	13.958598	NNP	1.750123	0.000570
ztnb	1	4	0.000031	14.958598	NNP	1.176091	0.000233
zwick	1	5	0.000031	14.958598	NNP	1.176091	0.000234
zz	5	2	0.000157	12.636670	NNP	1.176091	0.001170

4499 rows Ч 7 columns

In [ ]: