pdd-eda

September 8, 2021

0.0.1 Data exploring

Let's look at the data without any preprocesses steps, if we are lucky enough, we may find something interesting

[174]:

```
!pip install pandas
!pip install seaborn
!pip install matplotlib
!pip install wordcloud
!pip install nltk
```

```
Requirement already satisfied: pandas in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (1.2.3)
Requirement already satisfied: python-dateutil>=2.7.3 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from pandas)
(2.8.1)
Requirement already satisfied: pytz>=2017.3 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from pandas)
(2021.1)
Requirement already satisfied: numpy>=1.16.5 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from pandas)
(1.20.2)
Requirement already satisfied: six>=1.5 in
/Users/mbao01/.local/lib/python3.8/site-packages (from python-
dateutil>=2.7.3->pandas) (1.14.0)
Requirement already satisfied: seaborn in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (0.11.1)
Requirement already satisfied: scipy>=1.0 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from seaborn)
(1.6.2)
Requirement already satisfied: numpy>=1.15 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from seaborn)
(1.20.2)
Requirement already satisfied: matplotlib>=2.2 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from seaborn)
(3.3.4)
Requirement already satisfied: pandas>=0.23 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from seaborn)
(1.2.3)
```

```
Requirement already satisfied: kiwisolver>=1.0.1 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib>=2.2->seaborn) (1.3.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib>=2.2->seaborn) (2.4.7)
Requirement already satisfied: python-dateutil>=2.1 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib>=2.2->seaborn) (2.8.1)
Requirement already satisfied: pillow>=6.2.0 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib>=2.2->seaborn) (8.1.2)
Requirement already satisfied: cycler>=0.10 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib>=2.2->seaborn) (0.10.0)
Requirement already satisfied: six in /Users/mbao01/.local/lib/python3.8/site-
packages (from cycler>=0.10->matplotlib>=2.2->seaborn) (1.14.0)
Requirement already satisfied: pytz>=2017.3 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
pandas>=0.23->seaborn) (2021.1)
Requirement already satisfied: matplotlib in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (3.3.4)
Requirement already satisfied: numpy>=1.15 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from matplotlib)
(1.20.2)
Requirement already satisfied: pillow>=6.2.0 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from matplotlib)
(8.1.2)
Requirement already satisfied: cycler>=0.10 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from matplotlib)
(0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from matplotlib)
(1.3.1)
Requirement already satisfied: python-dateutil>=2.1 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from matplotlib)
(2.8.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from matplotlib)
(2.4.7)
Requirement already satisfied: six in /Users/mbao01/.local/lib/python3.8/site-
packages (from cycler>=0.10->matplotlib) (1.14.0)
Requirement already satisfied: wordcloud in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (1.8.1)
Requirement already satisfied: matplotlib in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from wordcloud)
(3.3.4)
Requirement already satisfied: numpy>=1.6.1 in
```

```
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from wordcloud)
(1.20.2)
Requirement already satisfied: pillow in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from wordcloud)
(8.1.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib->wordcloud) (2.4.7)
Requirement already satisfied: cycler>=0.10 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib->wordcloud) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib->wordcloud) (1.3.1)
Requirement already satisfied: python-dateutil>=2.1 in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from
matplotlib->wordcloud) (2.8.1)
Requirement already satisfied: six in /Users/mbao01/.local/lib/python3.8/site-
packages (from cycler>=0.10->matplotlib->wordcloud) (1.14.0)
Requirement already satisfied: nltk in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (3.5)
Requirement already satisfied: regex in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from nltk)
(2021.4.4)
Requirement already satisfied: click in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from nltk) (7.1.2)
Requirement already satisfied: joblib in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from nltk) (1.0.1)
Requirement already satisfied: tqdm in
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-packages (from nltk) (4.59.0)
```

[175]:

```python
import pandas as pd
import seaborn as sns, numpy as np
sns.set_theme(); np.random.seed(0)
import matplotlib.pyplot as plt
from wordcloud import WordCloud
import nltk
import urllib3.request
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA, TruncatedSVD
from sklearn.preprocessing import StandardScaler, MinMaxScaler
import numpy as np
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
# %matplotlib widget
```

[176]:

```python
df = pd.read_json("../data/dataset.v1.issues.json")
```

let's try select set of fields, that we think can help us and look at the report

0.0.2 First hypothesis

we have enough datapoints in labels field if so, the task can be solved as the classification one

[177]:

```
import json
with open("../data/dataset.v1.issues.json") as file:
    data = json.load(file)
```

[178]:

```
label_list = [issue['labels'] for issue in data['issues'] if 'labels' in issue]
label_dataset = []
for labels in label_list:
     label_dataset.append([label['name'] for label in labels if 'name' in␣
 ↪label])
df = pd.DataFrame(label_dataset)
```

0.0.3 Results of the investigating of first hypothesis

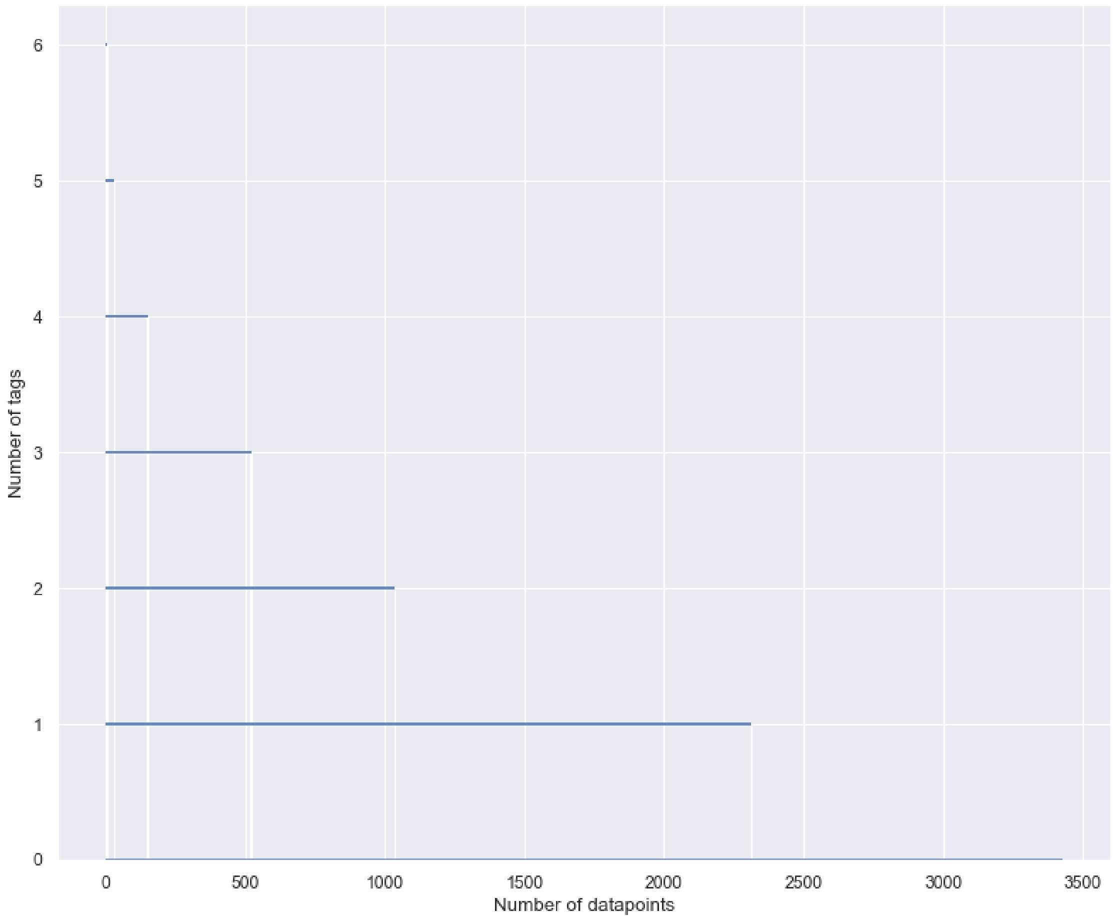
[179]:

```
print('1. Number of datapoints:', df.shape[0])
print('2. Number of datapoints with at least one tag:', df.shape[0] -  df[0].
 ↪isnull().sum())
print('3. Number of datapoints with at least 2 tags', df.shape[0] - df[1].
 ↪isnull().sum())
```

```
1. Number of datapoints: 5661
2. Number of datapoints with at least one tag: 3429
3. Number of datapoints with at least 2 tags 2311
```

[180]:

```
x = [df.shape[0] - df[i].isnull().sum() for i in range(7)]

plt.figure(figsize=(12, 10), dpi=80)
plt.bar(x,[i for i in range(7)],align='center')
plt.xlabel('Number of datapoints')
plt.ylabel('Number of tags')
for i in range(7):
    plt.hlines(i,0,x[i])
plt.show()
```

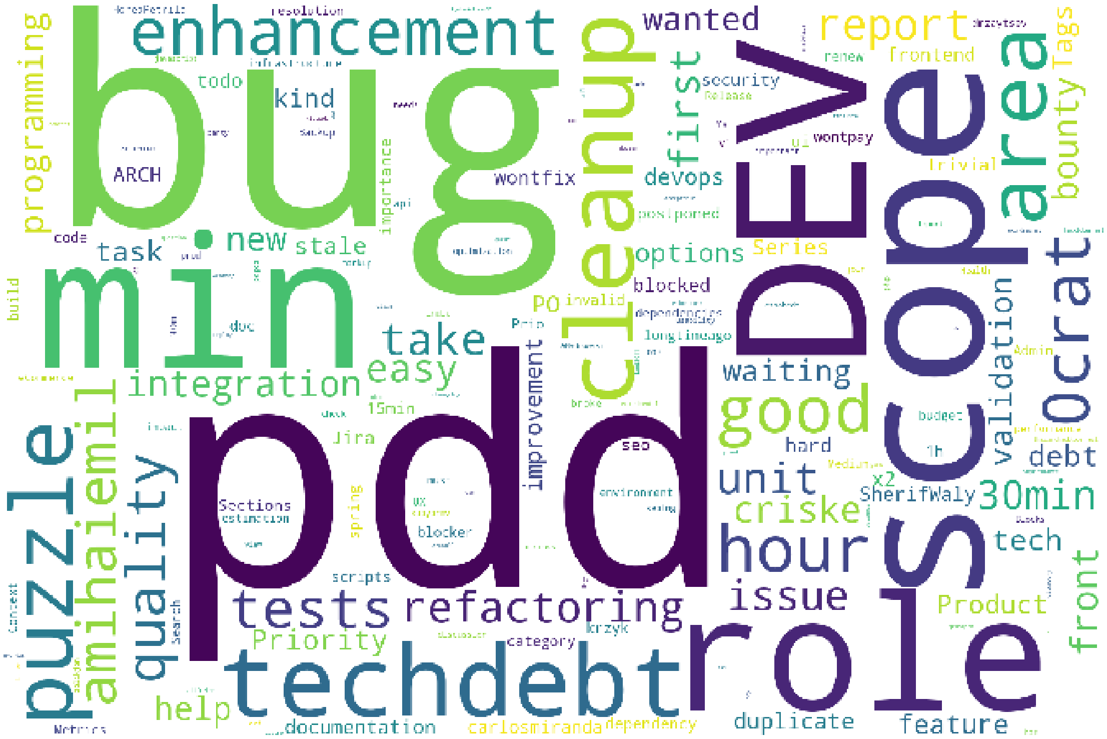


So we have pretty huge number of datapoints with at least one tag. But are they really helpful? I mean, is this label truly display severity of the puzzle?

[181]:
```
import itertools
```

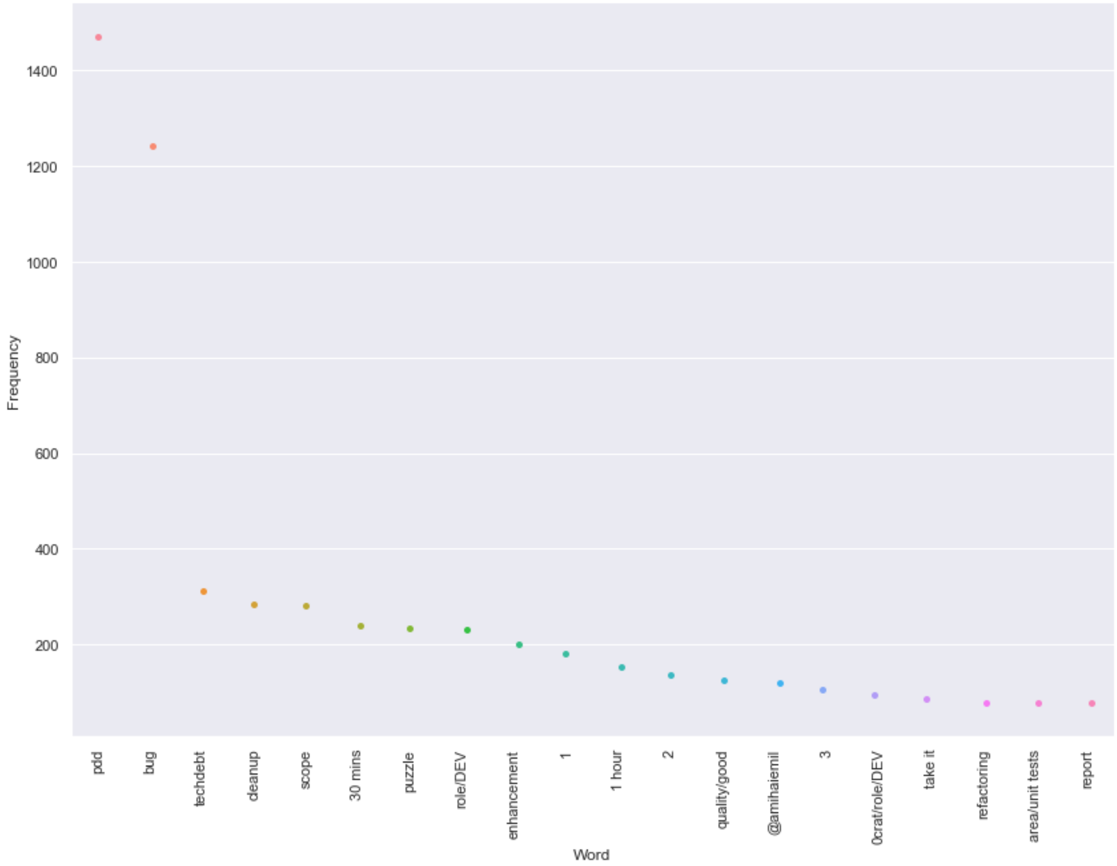
[182]:
```
tags = [tag for tag in list(itertools.chain(*[list(df[i]) for i in range(7)]))␣
 ↪if not pd.isnull(tag)]
tags_text = ' '.join(tags)
word_cloud = WordCloud(width=1800, height=1200,collocations = False,␣
 ↪background_color = 'white').generate(tags_text)
```

[183]:
```
plt.figure(figsize=(12, 10))
plt.imshow(word_cloud, interpolation='nearest')
plt.axis("off")
plt.show()
```



[184]:
```
word_dist = nltk.FreqDist(tags)
rslt=pd.DataFrame(word_dist.most_common(20),columns=['Word','Frequency'])
```

[185]:
```
sns.catplot(x='Word', y='Frequency',data = rslt,height=8.27, aspect=11.7/8.27);
plt.xticks(rotation=90);
```



0.0.4 From the observed data i can make several conclusion:

1. There are some good labels of puzzles, that actually contains information about severity of the bug, or the importance for the client, like: hour, easy, enhancement, first, documentation, techdebt.
2. Most of the puzzles contains labels, which is not very relevant to the ranking task, for example can we understand the importance of task with label bug or pdd?

0.0.5 As the conclusion for the first hypothesis, i propose to not to use the label field

0.0.6 Second hypothesis: if task is hard, many puzzles will be created, so if we have a long series of related puzzles, we can think of it as a one feature that need to be implemented. Longer the puzzle chain, then more importantly will be head of the chain

[186]:

```
df = pd.read_csv('../data/all-puzzles.csv')
df['Parent ID'].isnull().sum()
```

[186]:

```
10500
```

All puzzles do not have parent data, which means that we do not have data of relation of puzzles

0.0.7 At this moment, we can't use this approach

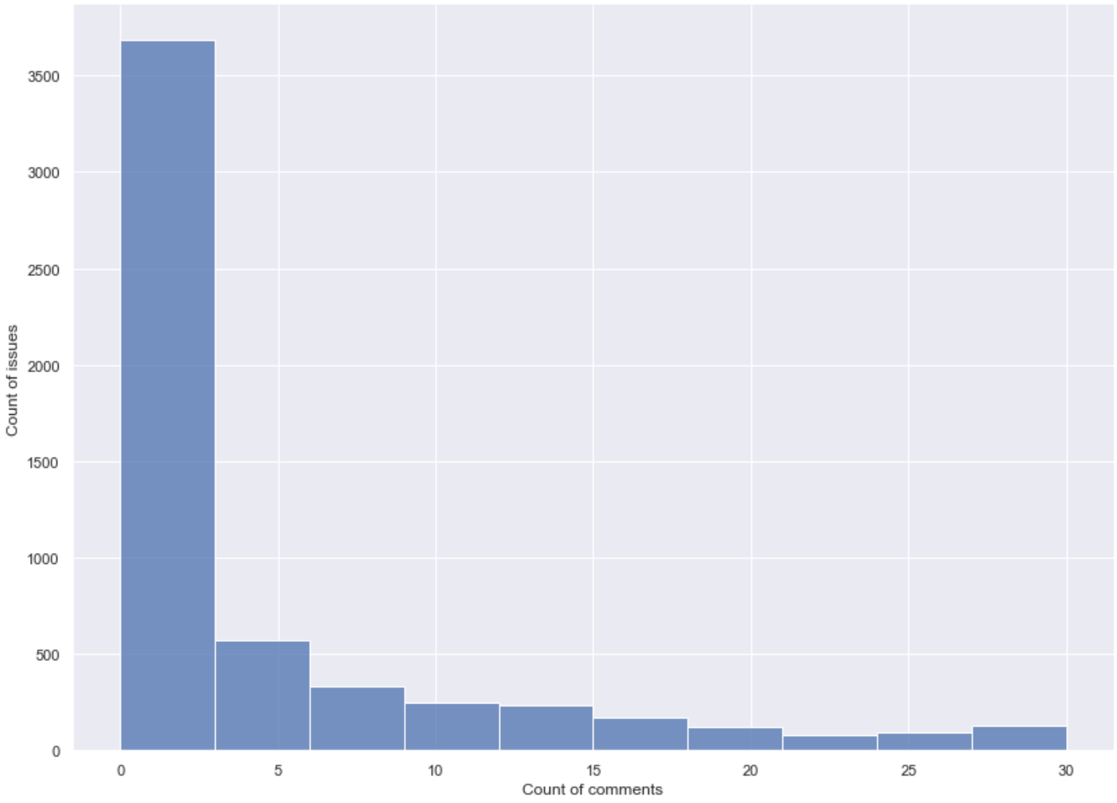
0.0.8 Another approach

If task is hard many people will be involved into this task, so we can predict the rating of the task, by predicting number of comments to the issue

```
[187]: with open('../data/dataset.v3.issues_extended_urls.json') as f:
           comments_count = []
           data = json.loads(f.read())
           data = [issue for issue in data['issues'] if 'comments_url' in issue]
           for issue in data:
               comments_count.append(len(issue['comments_url']))
```

```
[188]: ax = sns.displot(comments_count,binwidth=3, height=8.27, aspect=11.7/8.27)
       ax.set(xlabel='Count of comments', ylabel='Count of issues')
```

```
[188]: <seaborn.axisgrid.FacetGrid at 0x7f8981ffd880>
```



```
[189]: print('Count of issues where count of messages at least one: ',␣
        ↪len(list(filter(lambda x: x >= 1,comments_count))))
       print('Count of issues where count of messages is zero: ',␣
        ↪len(list(filter(lambda x: x == 0, comments_count))))
```

```
Count of issues where count of messages at least one:  4517
Count of issues where count of messages is zero:  1144
```

So even if we have very unbalanced data, where most of the issues have number of comments is less then 5, we could try this approach, when we predicting the rank of the puzzle, but predicting the count of comments

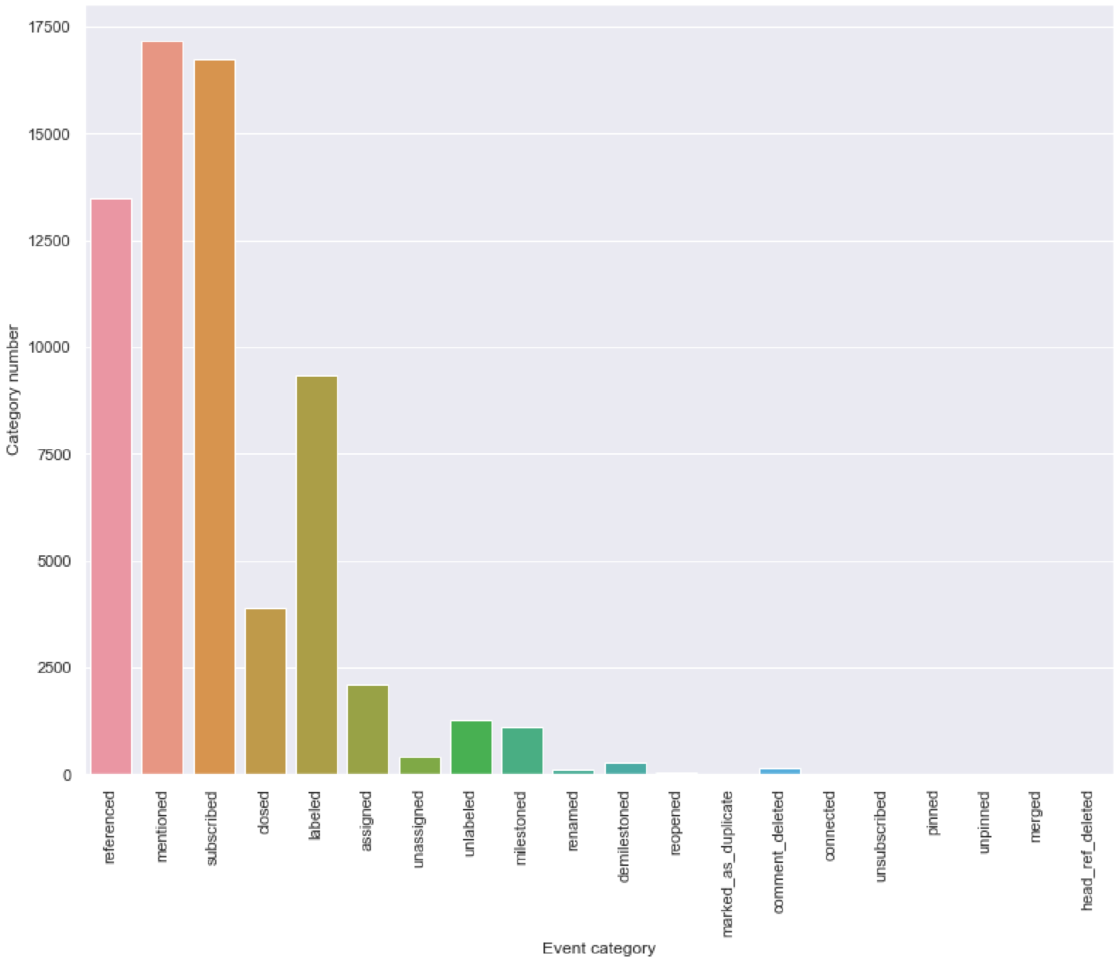
[190]:

```python
import json
import seaborn as sns
with open('../data/dataset.v3.issues_extended_urls.json') as f:
    events_type = []
    data = json.loads(f.read())
    data = [issue for issue in data['issues'] if 'events_url' in issue]
    for issue in data:
        events_type.extend([event['event'] for event in issue['events_url']])
```

[191]:

```python
import matplotlib.pyplot as plt
sns.set(rc={'figure.figsize':(13,10)})
ax = sns.countplot(events_type)#binwidth=3, height=8.27, aspect=11.7/8.27)
ax.set(xlabel='Event category', ylabel='Category number')
plt.xticks(rotation=90);
```

```
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning:

Pass the following variable as a keyword arg: x. From version 0.12, the only
valid positional argument will be `data`, and passing other arguments without an
explicit keyword will result in an error or misinterpretation.
```



[192]:

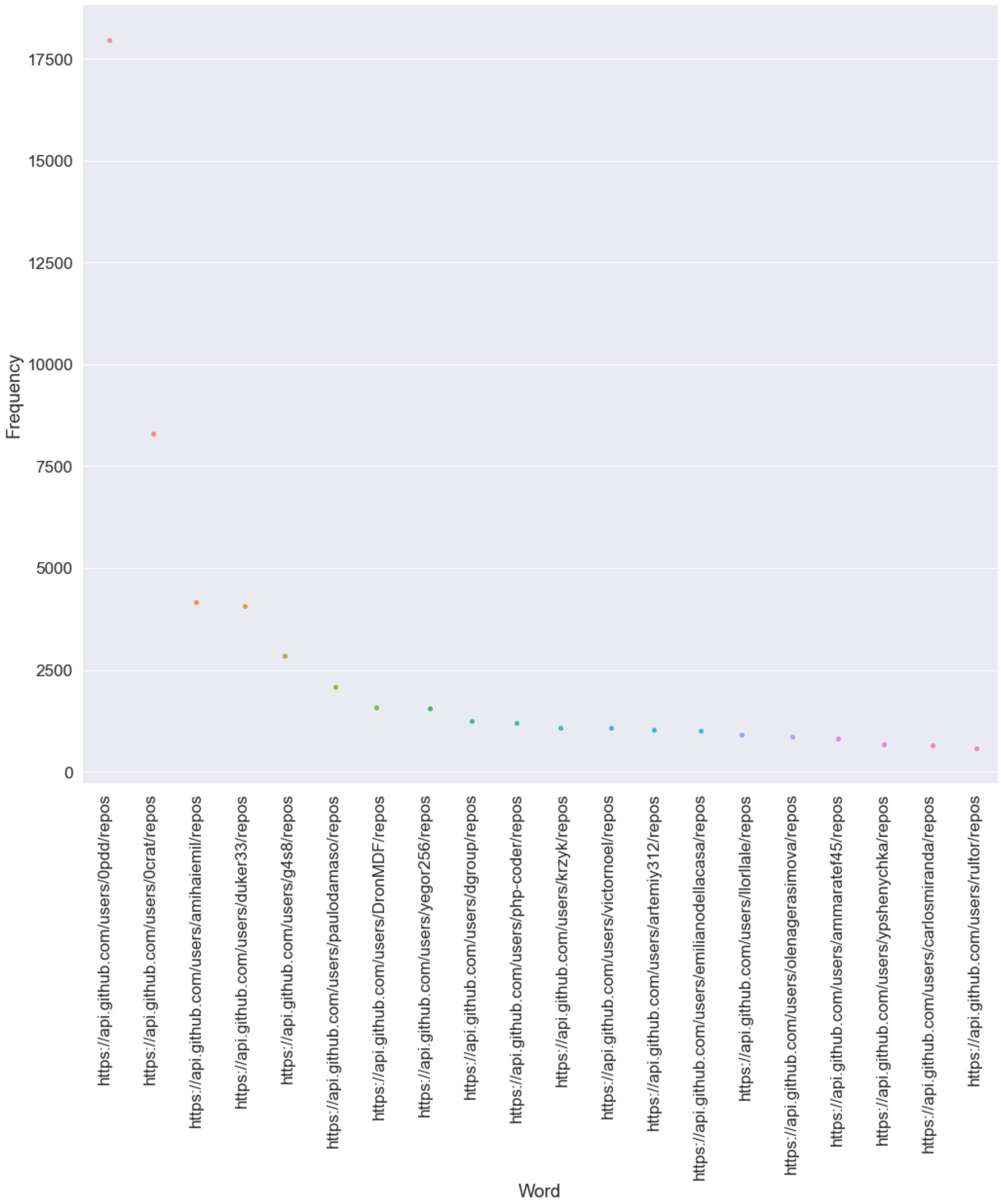
```
import json
import seaborn as sns
with open('../data/dataset.v3.issues_extended_urls.json') as f:
    actors_list = []
    data = json.loads(f.read())
    data = [issue for issue in data['issues'] if 'events_url' in issue]
    for issue in data:
        actors = ([event['actor'] for event in issue['events_url'] if 'actor'␣
 ↪in event and event['actor']])
        actors_list.extend([actor['repos_url'] for actor in actors if␣
 ↪'repos_url' in actor])
```

[193]:

```
sns.set()
sns.set(font_scale=1.55);
```

[194]:

```
import nltk
import pandas as pd
word_dist = nltk.FreqDist(actors_list)
rslt=pd.DataFrame(word_dist.most_common(20),columns=['Word','Frequency'])
sns.catplot(x='Word', y='Frequency',data = rslt,height=12, aspect=11.7/8.27);
plt.xticks(rotation=90);

# import matplotlib.pyplot as plt
# sns.set(rc={'figure.figsize':(13,10)})
# ax = sns.countplot(actors_list)#binwidth=3, height=8.27, aspect=11.7/8.27)
# ax.set(xlabel='Event category', ylabel='Category number')
# plt.xticks(rotation=90);
```



[195]:

```
import json
import seaborn as sns
from dateutil.parser import parse

fls = [f'../data/extended-issues-{i}.json' for i in range(3)]
```

```
def r(f):
    with open(f) as f:
        events_type = []
        data = json.loads(f.read())
    return data
```

[196]:

```
def get_row(issue):
    def _(f):
        return issue[f] if f in issue else None

    def _t(f1, f2):
        return (parse(_(f2)) - parse(_(f1))).total_seconds() if _(f1) and _(f2)␣
 ↪else None

    def __(f1, f2):
        return _(f1)[f2] if _(f1) and f2 in _(f1) else None

    def _l():
        return [label['name'] for label in _('labels')] if _('labels') else []

    time_delta = _t('created_at', 'closed_at')
    number_of_assignees = len(_('assignees') or [])
    changes = [__('code', 'additions') or 0,
               __('code', 'deletions') or 0
              ]

    comment_count = _('referenced') or 0
    mentioned = _('mentioned') or 0
    subscribed = _('subscribed') or 0
    number_of_files = len(__('all_commits','files') or [])
    labels = _l()
    return [time_delta, number_of_assignees, *changes, comment_count,␣
 ↪mentioned, subscribed,number_of_files, *labels]
```

[197]:

```
columns = ['time_delta', 'number_of_assignees', 'changes', 'comment_count',␣
 ↪'mentioned', 'subscribed','number_of_files', 'labels']
```

[230]:

```
overall_data = []
all_issues = []
for f in fls:
    data = r(f)
    all_issues.extend(data)
    overall_data.extend([get_row(issue) for issue in data if issue.get('state',␣
 ↪None) == 'closed'])
```

[231]:

```
print(overall_data[0])
closed_len = len([i for i in all_issues if i.get('state', None) == 'closed'])
```

```python
issues_len = len(all_issues)
open_len = len([i for i in all_issues if i.get('state', None) == 'open'])

print(f'No. of open issues {open_len}')
print(f'No. of closed issues {closed_len}')
print(f'Total no. of issues {issues_len}')
```

```
[103.0, 0, 2, 2, 2, 0, 0, 0]
No. of open issues 1509
No. of closed issues 4161
Total no. of issues 10489
```

[199]:

```python
pd_data = pd.DataFrame(overall_data)
```

[200]:

```python
import category_encoders as ce
encoder = ce.CountEncoder(normalize=True)
encoder.fit(np.array(list(itertools.chain(*pd_data.iloc[:,8:].to_numpy()))))
for i in range(8,15):
    pd_data.iloc[:,i] = encoder.transform(pd_data.iloc[:,i])[0]
```

[201]:

```python
scaler = MinMaxScaler()
scaler.fit(pd_data)
pd_data.iloc[:,:8] = scaler.fit_transform(pd_data.iloc[:,:8])
```

[202]:

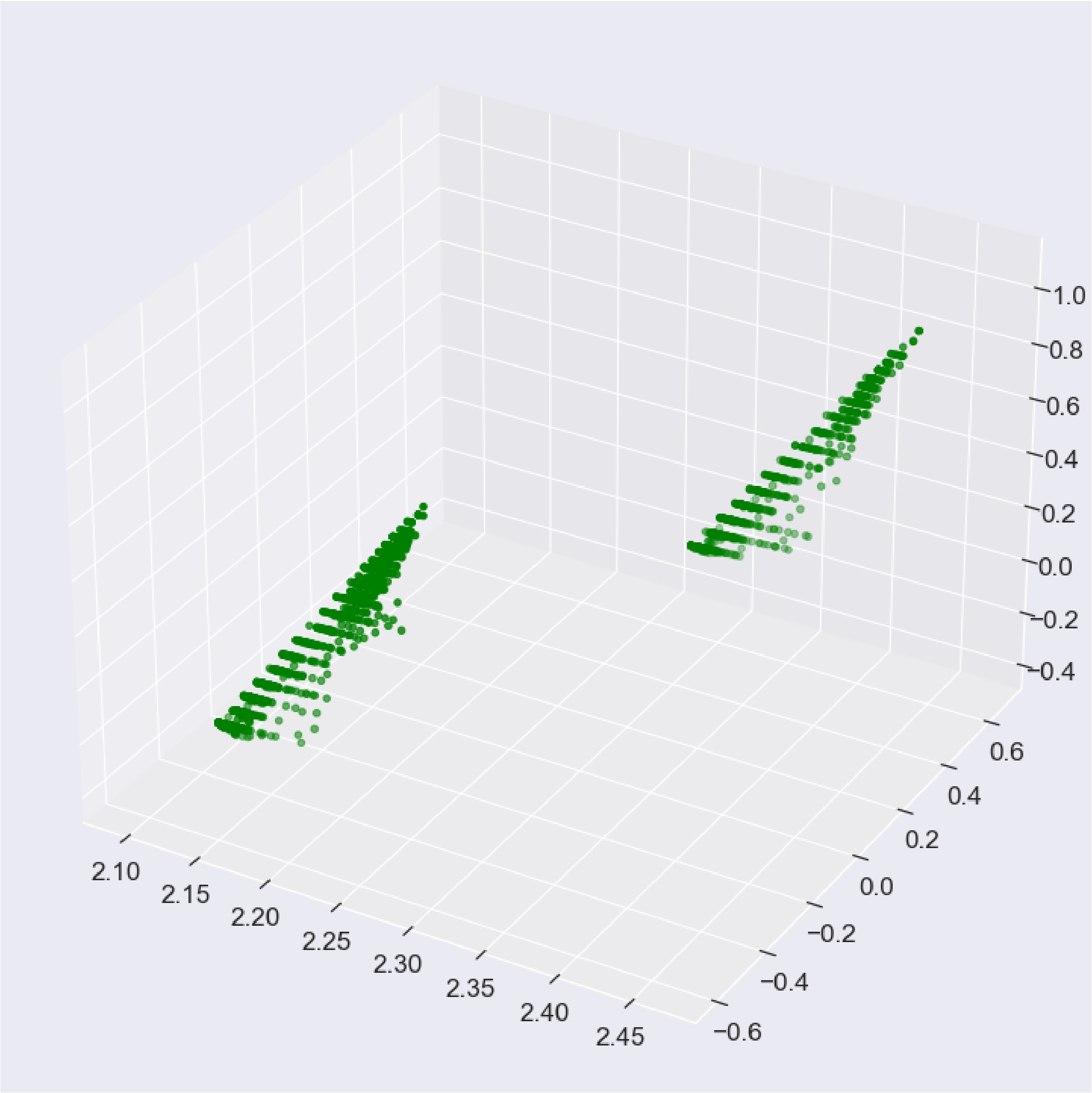
```python
pd_data = pd_data.dropna()
```

[203]:

```python
svd = TruncatedSVD(n_components=3, n_iter=10, random_state=42)
truncated = svd.fit_transform(pd_data)
```

[204]:

```python
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(truncated[:,0], truncated[:,1], truncated[:,2], color='green')
plt.show()
```



[205]:

```
import plotly
import plotly.graph_objs as go
plotly.offline.init_notebook_mode()

trace = go.Scatter3d(
    x= truncated[:,0],
    y= truncated[:,1],
    z= truncated[:,2],
    mode='markers',
    marker={
        'size': 10,
        'opacity': 0.8,
    }
)

layout = go.Layout(
    margin={'l': 0, 'r': 0, 'b': 0, 't': 0}
)

data = [trace]
plot_figure = go.Figure(data=data, layout=layout)
plotly.offline.iplot(plot_figure)
```

[206]:

```
svd.explained_variance_ratio_.sum()
```

[206]:

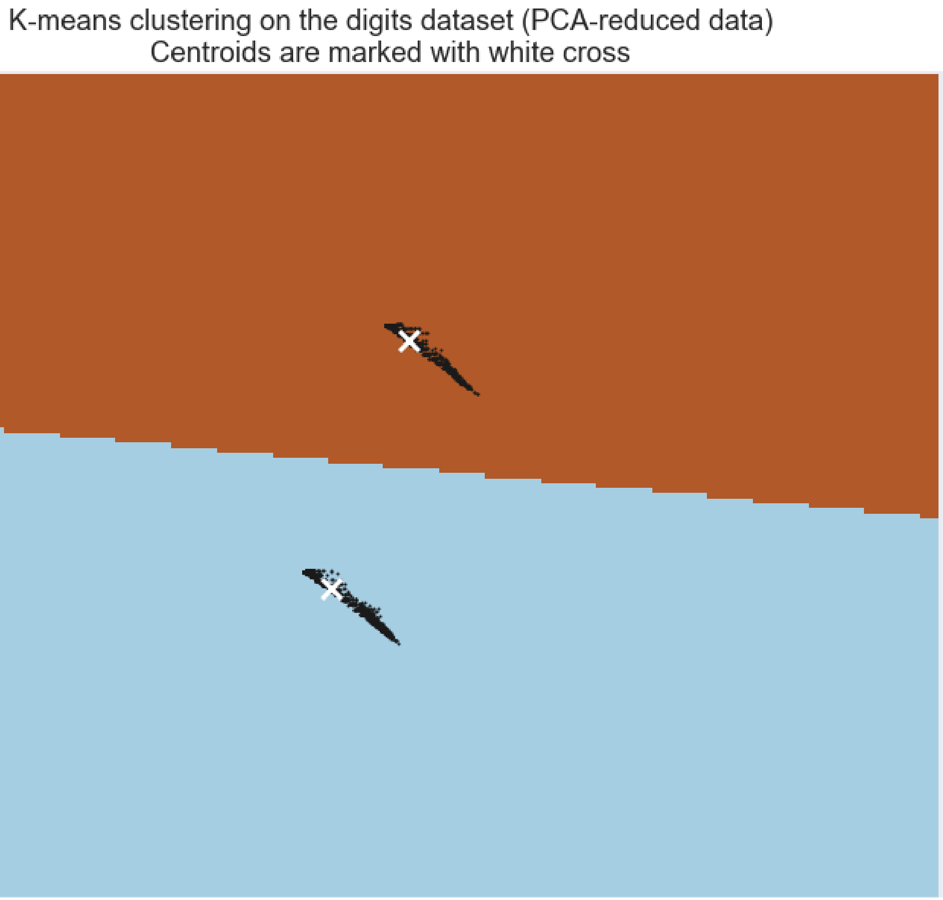
```
0.9374994440211427
```

[207]:

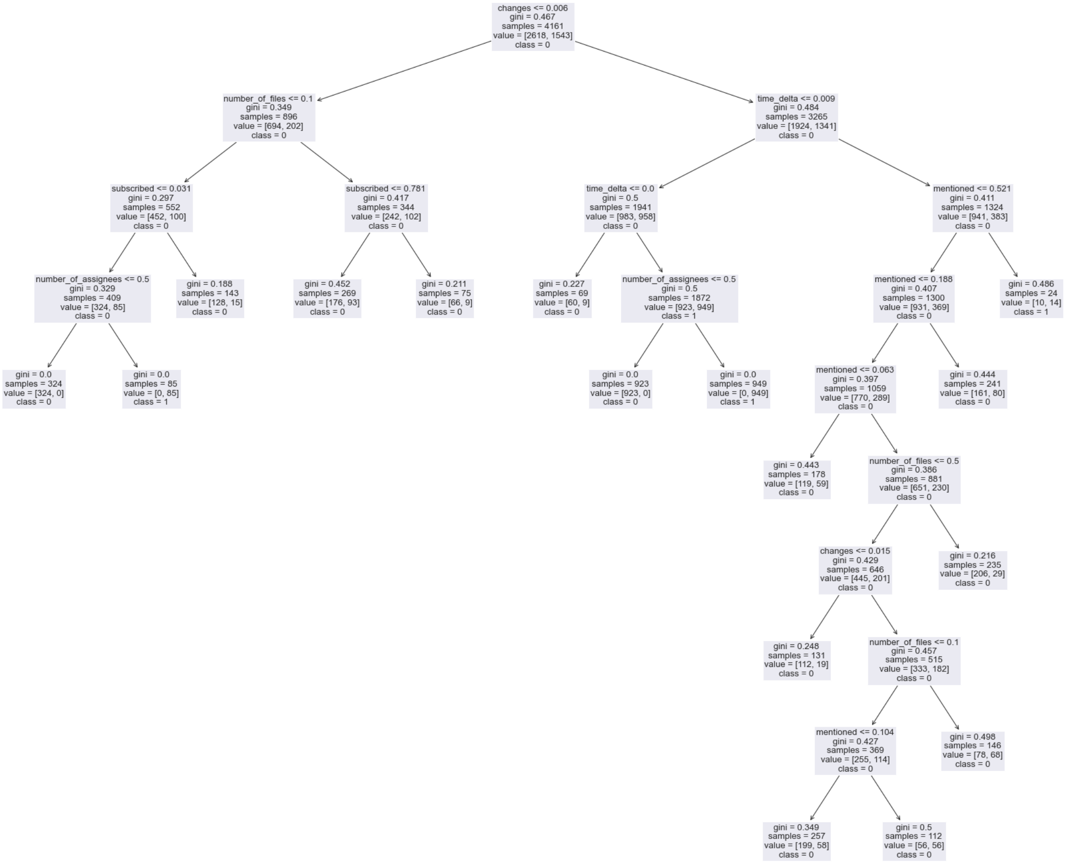
```
svd = TruncatedSVD(n_components=2, n_iter=10, random_state=42)
reduced_data = svd.fit_transform(pd_data)
X = np.array(reduced_data)
kmeans = KMeans(n_clusters=2, random_state=0).fit(X)
```

[208]:

```
h = .02

x_min, x_max = reduced_data[:, 0].min() - 1, reduced_data[:, 0].max() + 1
y_min, y_max = reduced_data[:, 1].min() - 1, reduced_data[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h))

Z = kmeans.predict(np.c_[xx.ravel(), yy.ravel()])

Z = Z.reshape(xx.shape)
plt.figure(1)
plt.clf()
plt.imshow(Z, interpolation="nearest",
           extent=(xx.min(), xx.max(), yy.min(), yy.max()),
           cmap=plt.cm.Paired, aspect="auto", origin="lower")

plt.plot(reduced_data[:, 0], reduced_data[:, 1], 'k.', markersize=2)
centroids = kmeans.cluster_centers_
plt.scatter(centroids[:, 0], centroids[:, 1], marker="x", s=169, linewidths=3,
            color="w", zorder=10)
plt.title("K-means clustering on the digits dataset (PCA-reduced data)\n"
          "Centroids are marked with white cross")
plt.xlim(x_min, x_max)
plt.ylim(y_min, y_max)
plt.xticks(())
plt.yticks(())
plt.show()
```



[209]:
```
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree

clf = DecisionTreeClassifier(random_state=42, max_features='log2',␣
 ↪min_samples_split=300,)
clf.fit(np.array(pd_data), kmeans.predict(X))
fig = plt.figure(figsize=(30,25))
tree.plot_tree(clf, feature_names=columns, class_names=['0', '1']);
```



[210]:

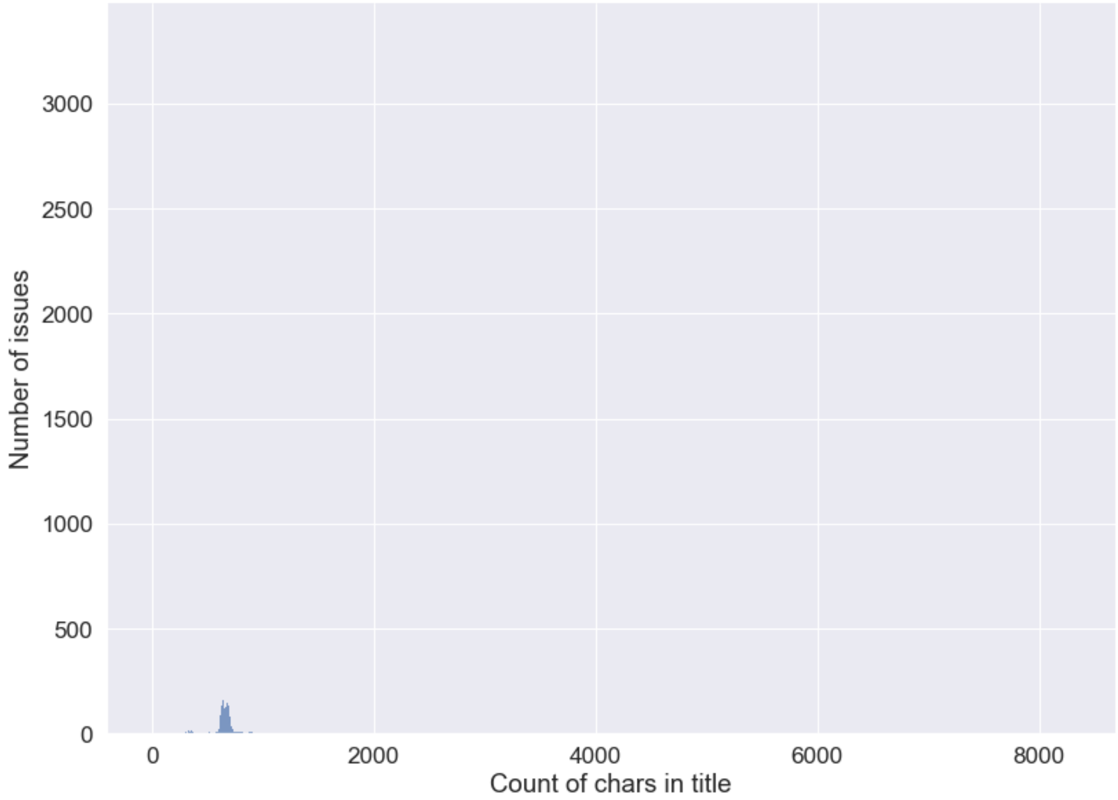
```
with open('../data/dataset.v1.issues.json') as f:
    data = json.loads(f.read())
    data = [issue.get('title', '').lower() + ' ' + issue.get('body', '').
 ↪lower() for issue in data['issues']]
```

[211]:

```
ax = sns.displot([len(issue) for issue in data],binwidth=3, height=8.27,␣
 ↪aspect=11.7/8.27)
ax.set(xlabel='Count of chars in title', ylabel='Number of issues')
```

[211]:

```
<seaborn.axisgrid.FacetGrid at 0x7f897f6f1d00>
```

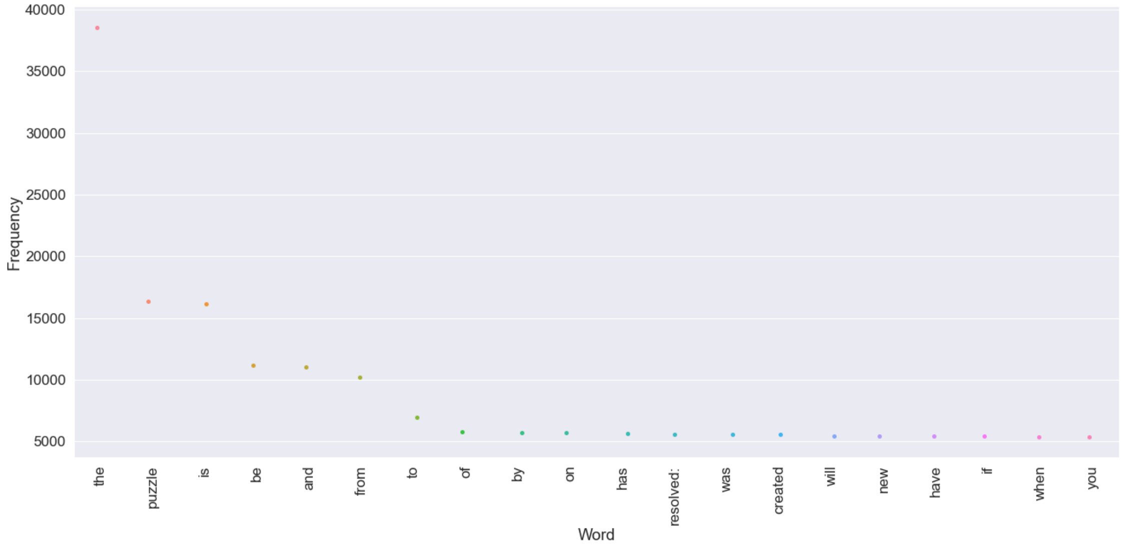


```
[212]: word_dist = nltk.FreqDist(data)
       rslt=pd.DataFrame(word_dist.most_common(20),columns=['Word','Frequency'])
       sns.catplot(x='Word', y='Frequency',data = rslt,height=8.27, aspect=11.7/5.27);
       plt.xticks(rotation=90);
```

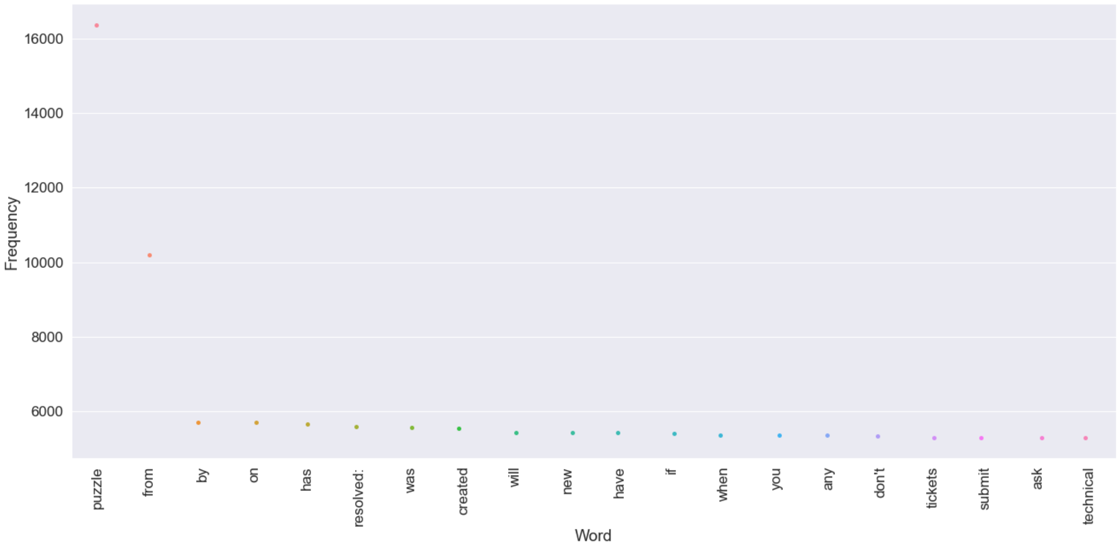
```
/Users/mbao01/.conda/envs/py3.8/lib/python3.8/site-
packages/seaborn/axisgrid.py:64: UserWarning:

Tight layout not applied. The left and right margins cannot be made large enough
to accommodate all axes decorations.
```

Frequency
0
500
1000
1500
2000
2500
3000
Word
001-install-main-lazylead-tables.sql:69: properties.type... the puzzle `dev-347aeee3` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/dc6a60e92d33925262186c94230dc9fde54b134d/upgrades/sqlite/001-install-main-lazylead-tables.sql#l69-l69
the puzzle was created by @dgroup on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
docker-compose.yml:3: define openshift template in order... the puzzle `dev-dd5a7571` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/dc6a60e92d33925262186c94230dc9fde54b134d/.docker/docker-compose.yml#l3-l3
the puzzle was created by @dgroup on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
docker-compose.yml:14: the logs(including gc) should be... the puzzle `dev-aeb32344` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/dc6a60e92d33925262186c94230dc9fde54b134d/.docker/docker-compose.yml#l14-l14
the puzzle was created by @dgroup on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
todo.yml:1-2: enable travis ci as second ci server once... the puzzle `dev-1214b7cd` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/dc6a60e92d33925262186c94230dc9fde54b134d/todo.yml#l1-l2
the puzzle was created by yurii dubinka on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
todo.yml:4-5: explain in readme.md how to use the... the puzzle `dev-7f4f0ad5` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/dc6a60e92d33925262186c94230dc9fde54b134d/todo.yml#l4-l5
the puzzle was created by yurii dubinka on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
todo.yml:7-10: there a lot of badges in readme which... the puzzle `dev-6c09bce3` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/dc6a60e92d33925262186c94230dc9fde54b134d/todo.yml#l7-l10
the puzzle was created by yurii dubinka on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
todo.yml:12-13: migrate to ruby 2.7.0, for now the... the puzzle `dev-dab00982` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/dc6a60e92d33925262186c94230dc9fde54b134d/todo.yml#l12-l13
the puzzle was created by yurii dubinka on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
todo.yml:15-16: enable depshield badge for inspection... the puzzle `dev-6347430d` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/dc6a60e92d33925262186c94230dc9fde54b134d/todo.yml#l15-l16
the puzzle was created by yurii dubinka on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
001-install-main-lazylead-tables.sql:68: task.enabled -... the puzzle `dev-9c14b40d` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/1691dfbf7e1706b7d20b33bb24b531734fbfadd3/upgrades/sqlite/001-install-main-lazylead-tables.sql#l68-l68
the puzzle was created by @dgroup on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
rakefile:31-37: investigate the possibility of using... the puzzle `dev-474a56c9` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/c6d6bb638b97c748ca88d6987b52fc8914f22515/rakefile#l31-l37
the puzzle was created by yurii dubinka on 19-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
001-install-main-lazylead-tables.sql:24-25: add indexes... the puzzle `dev-2823e14f` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/bffa62405c1a5e5e740926f861e323d9151c785f/upgrades/sqlite/001-install-main-lazylead-tables.sql#l24-l25
the puzzle was created by @dgroup on 26-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
001-install-main-lazylead-tables.sql:30-31: add index by... the puzzle `dev-54a63de1` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/bffa62405c1a5e5e740926f861e323d9151c785f/upgrades/sqlite/001-install-main-lazylead-tables.sql#l30-l31
the puzzle was created by @dgroup on 26-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
001-install-main-lazylead-tables.sql:73: tasks.enabled -... the puzzle `dev-7d63e4e1` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/bffa62405c1a5e5e740926f861e323d9151c785f/upgrades/sqlite/001-install-main-lazylead-tables.sql#l73-l73
the puzzle was created by @dgroup on 26-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
lazylead.gemspec:30-31: use single approach for defining... the puzzle `dev-6ce436fe` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/bffa62405c1a5e5e740926f861e323d9151c785f/lazylead.gemspec#l30-l31
the puzzle was created by @dgroup on 26-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
model.rb:30-31: setup the relations between classes... the puzzle `dev-988fcfd0` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/bffa62405c1a5e5e740926f861e323d9151c785f/lib/lazylead/orm/model.rb#l30-l31
the puzzle was created by yurii dubinka on 26-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
model.rb:33-34: add validations to the columns, more... the puzzle `dev-7c4e32db` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/bffa62405c1a5e5e740926f861e323d9151c785f/lib/lazylead/orm/model.rb#l33-l34
the puzzle was created by yurii dubinka on 26-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
start_test.rb:74-81: think about using "timecop" >v0.9.1... the puzzle `10-74e00feb` from #10 has to be resolved:
https://github.com/dgroup/rll/blob/b699bd31b78aeac4e65db80eb78ed68b59ef3a5c/test/lazylead/cli/start_test.rb#l74-l81
the puzzle was created by yurii dubinka on 28-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
001-install-main-lazylead-tables.sql:68-71: task.cron -... the puzzle `dev-6508c131` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/b699bd31b78aeac4e65db80eb78ed68b59ef3a5c/upgrades/sqlite/001-install-main-lazylead-tables.sql#l68-l71
the puzzle was created by @dgroup on 28-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).
.rubocop.yml:74-75: add violation regarding methods... the puzzle `dev-a80e719e` from #dev has to be resolved:
https://github.com/dgroup/rll/blob/9a9f7b520f72de4b1881c3589d0ba761f17b4103/.rubocop.yml#l74-l75
the puzzle was created by yurii dubinka on 28-jan-20.
role: dev.
if you have any technical questions, don't ask me, submit new tickets instead. the task will be \"done\" when the problem is fixed and the text of the puzzle is _removed_ from the source code. here is more about [pdd](http://www.yegor256.com/2009/03/04/pdd.html) and [about me](http://www.yegor256.com/2017/04/05/pdd-in-action.html).

[213]:
```
words = [word for sentence in data for word in sentence.split()]
word_dist = nltk.FreqDist(words)
rslt=pd.DataFrame(word_dist.most_common(20),columns=['Word','Frequency'])
sns.catplot(x='Word', y='Frequency',data = rslt,height=8.27, aspect=11.7/5.27);
plt.xticks(rotation=90);
```



[214]:
```
stop_words = ['to','the','for', 'this','a', 'and','of','is','we','be','pdd']

words = [word for word in words if word not in stop_words]
word_dist = nltk.FreqDist(words)
rslt=pd.DataFrame(word_dist.most_common(20),columns=['Word','Frequency'])
sns.catplot(x='Word', y='Frequency',data = rslt,height=8.27, aspect=11.7/5.27);
plt.xticks(rotation=90);
```



[215]:

```
from nltk.tokenize import word_tokenize
nltk.download('punkt')

df = pd.DataFrame({'issues': data })
df['cleaned'] = df['issues'].str.replace("[^a-zA-Z#]", " ")
df['cleaned'] = df['cleaned'].apply(lambda x: np.nan if x.strip() == '' else x)
df = df[df['cleaned'].notnull()]

df = df['cleaned'].apply(lambda x: [word for word in word_tokenize(x) if␣
 ↪len(word)>1] if x else [''])
df = df.apply(lambda x: [item for item in x if item not in stop_words])

detokenized_doc = []
for i in range(len(df)):
    if i in df:
        t = ' '.join(df[i])
        detokenized_doc.append(t)
```

```
[nltk_data] Downloading package punkt to /Users/mbao01/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
<ipython-input-215-17c823b29e13>:5: FutureWarning:

The default value of regex will change from True to False in a future version.
```

[216]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
# tfidf vectorizer of scikit learn
vectorizer = TfidfVectorizer(max_features= 500, ngram_range=(1,3))
X = vectorizer.fit_transform(detokenized_doc)
```

```
terms = vectorizer.get_feature_names()
```

[217]:

```
svd = TruncatedSVD(n_components=3, n_iter=10, random_state=42)
truncated = svd.fit_transform(X)

trace = go.Scatter3d(
    x= truncated[:,0],
    y= truncated[:,1],
    z= truncated[:,2],
    mode='markers',
    marker={
        'size': 10,
        'opacity': 0.8,
    }
)

layout = go.Layout(
    margin={'l': 0, 'r': 0, 'b': 0, 't': 0}
)

data = [trace]
plot_figure = go.Figure(data=data, layout=layout)
plotly.offline.iplot(plot_figure)
```

[218]:

```
svd.explained_variance_ratio_.sum()
```

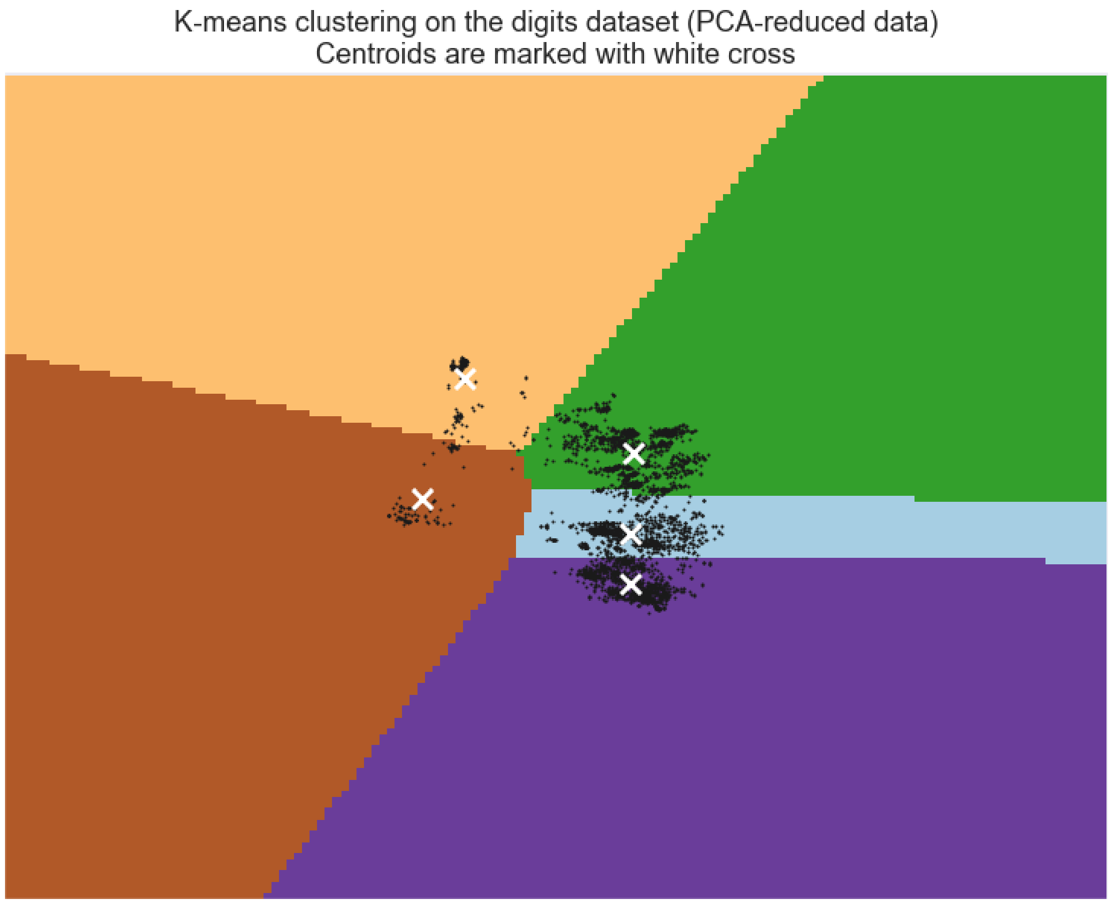
[218]:

```
0.16487977935494508
```

[219]:

```
svd = TruncatedSVD(n_components=2, n_iter=15, random_state=42)
reduced_data = svd.fit_transform(X)
X = np.array(reduced_data)
kmeans = KMeans(n_clusters=5, random_state=0).fit(X)
h = .02

x_min, x_max = reduced_data[:, 0].min() - 1, reduced_data[:, 0].max() + 1
y_min, y_max = reduced_data[:, 1].min() - 1, reduced_data[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h))

Z = kmeans.predict(np.c_[xx.ravel(), yy.ravel()])

Z = Z.reshape(xx.shape)
plt.figure(1)
plt.clf()
plt.imshow(Z, interpolation="nearest",
           extent=(xx.min(), xx.max(), yy.min(), yy.max()),
           cmap=plt.cm.Paired, aspect="auto", origin="lower")
```

```
plt.plot(reduced_data[:, 0], reduced_data[:, 1], 'k.', markersize=2)
centroids = kmeans.cluster_centers_
plt.scatter(centroids[:, 0], centroids[:, 1], marker="x", s=169, linewidths=3,
            color="w", zorder=10)
plt.title("K-means clustering on the digits dataset (PCA-reduced data)\n"
          "Centroids are marked with white cross")
plt.xlim(x_min, x_max)
plt.ylim(y_min, y_max)
plt.xticks(())
plt.yticks(())
plt.show()
```



[220]:

```
svd.explained_variance_ratio_.sum()
```

[220]: 0.10749240243147687

[221]:

```
from sklearn.utils.extmath import randomized_svd
U, Sigma, VT = randomized_svd(X, n_components=10, n_iter=100,
                              random_state=122)
```

```
#printing the concepts
for i, comp in enumerate(VT):
        terms_comp = zip(terms, comp)
        sorted_terms = sorted(terms_comp, key= lambda x:x[1], reverse=True)[:7]
        print("Concept "+str(i)+": ")
        for t in sorted_terms:
            print(t[0])
        print(" ")
```

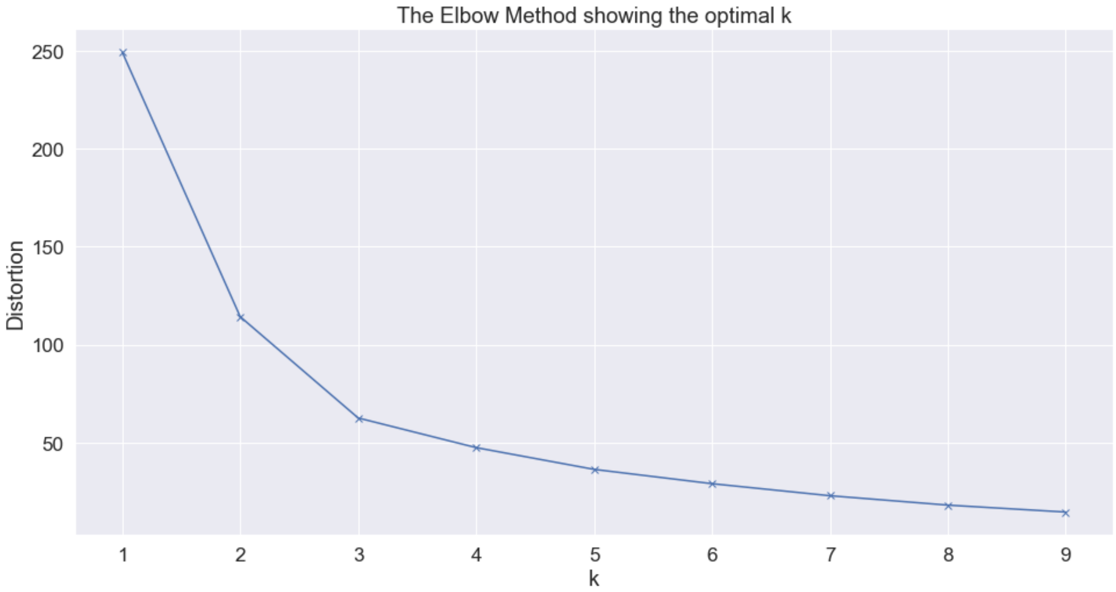
```
Concept 0:
aa
ab

Concept 1:
ab
aa
```

[222]:

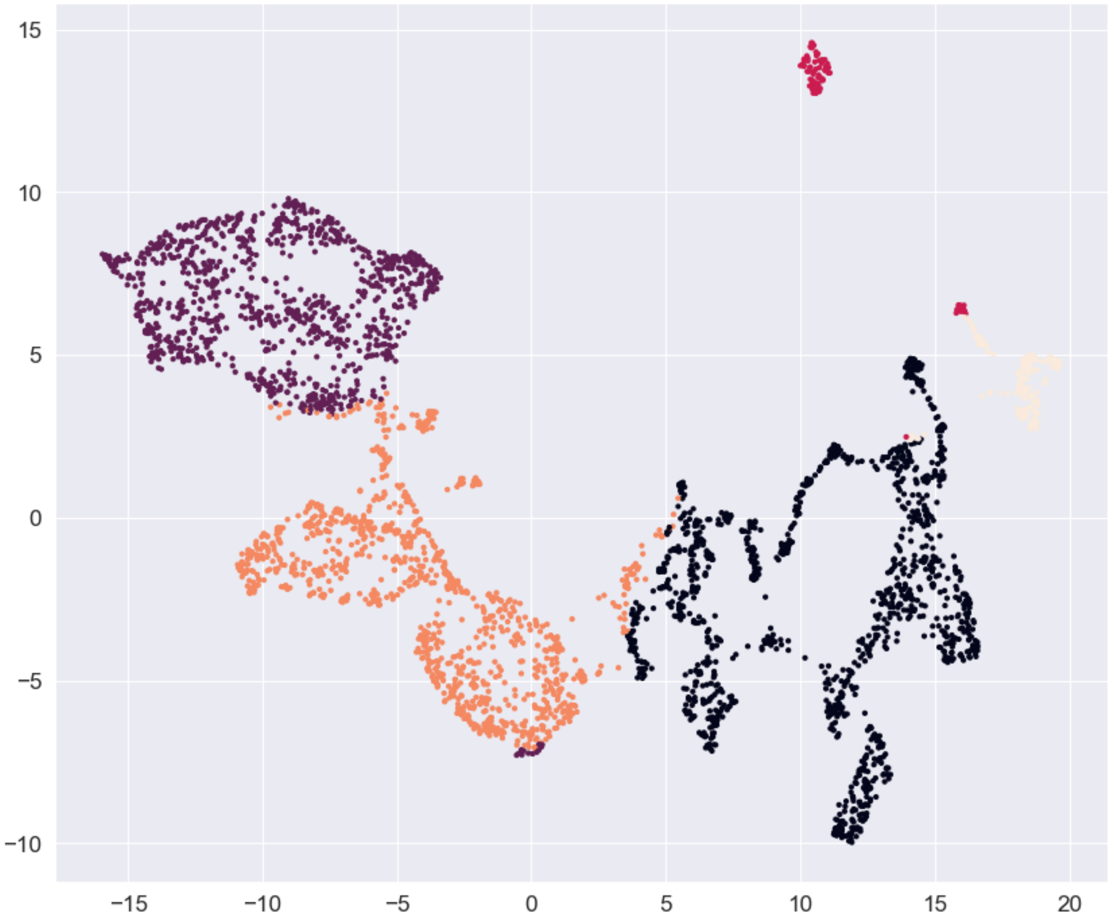
```
import umap.umap_ as umap
X_topics=U*Sigma
embedding = umap.UMAP(n_neighbors=100, min_dist=0.5, random_state=12).
 ↪fit_transform(X_topics)
```

[223]:

```
distortions = []
K = range(1,10)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(X)
    distortions.append(kmeanModel.inertia_)
plt.figure(figsize=(16,8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```



[224]:
```
from sklearn.cluster import KMeans
num_clusters = 5
km = KMeans(n_clusters=num_clusters)
km.fit(X)
clusters = km.labels_.tolist()
plt.figure(figsize=(14,12))
plt.scatter(embedding[:, 0], embedding[:, 1], c = clusters,s = 10,)
plt.show()
```


15
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
20