Analise the semeval2019 data

In [1]:
```
from emotion_recognition.helpers_explore_data import *
from emotion_recognition.starting_kit.baseline import preprocessdata

import pandas as pd

from pylab import rcParams
```
```
Using TensorFlow backend.
```

train set

In [2]:
```
train = pd.read_csv('../data/semeval2019/starterkitdata/train.txt', sep='\t')
train_text = list(train["turn1"] + train["turn2"] + train["turn3"])
_, train_preprocessed, train_labels = preprocessdata('../data/semeval2019/starterkitdata/train.txt', 'train', eos=
```
```
...
```

In [3]:
```
train.head(10)
```

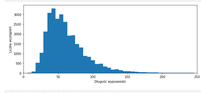
Out[3]:

	id	turn1	turn2	turn3	label
0	0	Don't worry I'm girl	hmm how do I know if you are	What's ur name?	others
1	1	When did IT	see many times i think ...	No, I never see you	angry
2	2	By	by Google Chrome	Where you live	others
3	3	U r ridiculous	I might be ridiculous but I am telling the truth.	U little disgusting whore	angry
4	4	Just for time pass	wt do u do 4 a living then	Maybe	others
5	5	I'm a dog person	yours so rude	Whatsaid why	others
6	6	So whatsup	Nothing much. Sitting sipping and watching TV....	What are you watching on tv?	others
7	7	Ok	ok im back?	So, how are u	others
8	8	Really?	really really really really really ...	Y saying so many times, .. i can hear you	others
9	9	Bay	in the bay	I see you	others

In [4]:
```
train['label'].describe()
```

Out[4]:
```
count     30160
unique        4
top      others
freq      14948
Name: label, dtype: object
```

In [5]:
```
rcParams['figure.figsize'] = 10, 4
plt.hist([len(s) for s in train_text], 120)
plt.xlabel('Długość wypowiedzi')
plt.ylabel('Liczba wystąpień')
plt.xlim(0, 250)
plt.show()
```



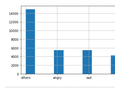
In [6]:
```
pd.DataFrame([len(s) for s in train_text]).describe()
```

Out[6]:

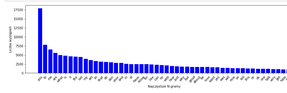
	0
count	30160.000000
mean	62.280305
std	38.189823
min	6.000000
25%	42.000000
50%	56.000000
75%	75.000000
max	692.000000

In [7]:
```
rcParams['figure.figsize'] = 6, 4
train['label'].hist()
```

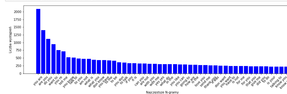
Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x14013b090>
```



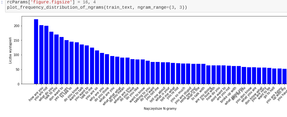
In [8]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(train_text, ngram_range=(1, 1))
```



In [9]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(train_text, ngram_range=(2, 2))
```



In [10]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(train_text, ngram_range=(3, 3))
```



dev set

In [11]:
```
dev = pd.read_csv('../data/semeval2019/devsetwithlabels/dev.txt', sep='\t')
dev_text = list(dev["turn1"] + dev["turn2"] + dev["turn3"])
```

In [12]:
```
dev.head(10)
```

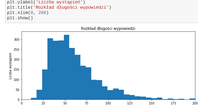
Out[12]:

	id	turn1	turn2	turn3	label
0	0	Then dont ask me	YOURE A GUY NOT AS IF YOU WOULD UNDERSTAND	IM NOT A GUY FUCK OFF	angry
1	1	Mixed things such as??	the things you do.	Have you seen romans??	others
2	2	Today I'm very happy	and I'm happy for you ❤	I will be many	happy
3	3	Woah bring me some	left it there now.	Brb	others
4	4	it is thuwooo	I said soon master.	he is pressuring me	others
5	5	Wont u ask my age??	Hey at least I sign well	Can u tell me how can we get close??	others
6	6	I said yes	What if I tell you I'm not?	Go to hell	angry
7	7	Where I'll check	why tomorrow?	No I want now	others
8	8	What we meet	you say: you're leaving soon...anywhere you wa...	?	others
9	9	Lets change the subject	I just did it :(	You're broken	sad

In [13]:
```
dev['label'].describe()
```

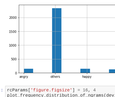
Out[13]:
```
count     2755
unique       4
top     others
freq      2338
Name: label, dtype: object
```

In [14]:
```
rcParams['figure.figsize'] = 10, 4
plt.hist([len(s) for s in dev_text], 50)
plt.xlabel('Długość wypowiedzi')
plt.ylabel('Liczba wystąpień')
plt.title('Rozkład długości wypowiedzi')
plt.xlim(0, 200)
plt.show()
```

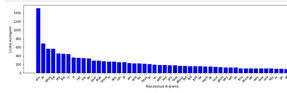


In [15]:
```
rcParams['figure.figsize'] = 6, 4
dev['label'].hist()
```

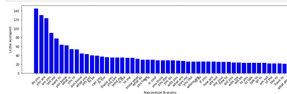
Out[15]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x14d22e588>
```



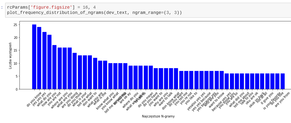
In [16]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(dev_text, ngram_range=(1, 1))
```



In [17]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(dev_text, ngram_range=(2, 2))
```



In [18]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(dev_text, ngram_range=(3, 3))
```



test set

In [19]:
```
test = pd.read_csv('../data/semeval2019/test/test.txt', sep='\t')
test_text = list(test["turn1"] + test["turn2"] + test["turn3"])
```

In [20]:
```
test.head(15)
```

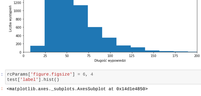
Out[20]:

	id	turn1	turn2	turn3	label
0	0	Hmm	What does your like mean?	I don't have any bra	others
1	1	What you like	very little things	Ok	others
2	2	Yes.	How so?	I want to fuck baby.	others
3	3	what did you guess	what what	fuck	others
4	4	We ?	of course we will!	What gender movies you like??	others
5	5	Where are you now?	At home just about to have breakfast...	what are you eating?	others
6	6	That was a joke btw...	It was	Yes :)	happy
7	7	Who d hell s he	johnny depp...duh?	Who she	others
8	8	yes, good advice	best advice ...	i great the	others
9	9	Nice to meet u	Hi, nice to meet you too!	:)	happy
10	10	Yuqq	why?	Don't know I'm tired	others
11	11	Software	Software what? i plan on going for development...	i am into android development stuff	others
12	12	Very nice	Thanks!! :)	R u know tamil	others
13	13	First you for me	Okay	So I talked rude	angry
14	14	Love you	you don't recognize me?	I love you ::	others

In [21]:
```
test['label'].describe()
```

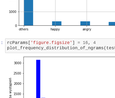
Out[21]:
```
count     5509
unique       4
top     others
freq      4077
Name: label, dtype: object
```

In [22]:
```
rcParams['figure.figsize'] = 10, 4
plt.hist([len(s) for s in test_text], 50)
plt.xlabel('Długość wypowiedzi')
plt.ylabel('Liczba wystąpień')
plt.title('Rozkład długości wypowiedzi')
plt.xlim(0, 200)
plt.show()
```

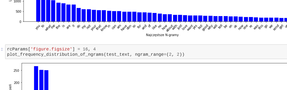


In [23]:
```
rcParams['figure.figsize'] = 6, 4
test['label'].hist()
```

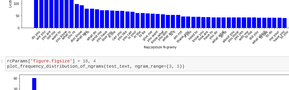
Out[23]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x145e4e898>
```



In [24]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(test_text, ngram_range=(1, 1))
```



In [25]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(test_text, ngram_range=(2, 2))
```



In [26]:
```
rcParams['figure.figsize'] = 10, 4
plot_frequency_distribution_of_ngrams(test_text, ngram_range=(3, 3))
```