NLP - lab1

Mateusz Praski

In [1]:
```
import regex
import matplotlib

import matplotlib.pyplot as plt
import pandas as pd
```

In [2]:
```
matplotlib.rcParams['figure.figsize'] = (16, 7)
```

In [3]:
```
df = pd.read_json("../../data/corpus.jsonl", lines=True)
df = df.set_index('_id').sort_index()
df.head()
```

Out[3]:

_id	title	text	metadata
3		Nie mówię, że nie podoba mi się też pomysł szk...	{}
31		Tak więc nic nie zapobiega fałszywym ocenom po...	{}
56		Nigdy nie możesz korzystać z FSA dla indywidua...	{}
59		Samsung stworzył LCD i inne technologie płaski...	{}
63		Oto wymagania SEC: Federalne przepisy dotycząc...	{}

In [4]:
```
corpus = df['text']
corpus[3]
```

Out[4]:
```
'Nie mówię, że nie podoba mi się też pomysł szkolenia w miejscu pracy, ale
nie możesz oczekiwać, że firma to zrobi. Szkolenie pracowników to nie ich p
raca — oni tworzą oprogramowanie. Być może systemy edukacyjne w Stanach Zje
dnoczonych (lub ich studenci) powinny trochę martwić się o zdobycie umiejęt
ności rynkowych w zamian za ich ogromne inwestycje w edukację, zamiast wych
odzić z tysiącami zadłużonych studentów i narzekać, że nie są do niczego wy
kwalifikowani.'
```

In [5]:
```
time_regex = r"\b(?:(?:2[0-4])|(?:1?\d)):(?:[0-5]\d)(?::[0-5]\d)?"

months_regex = [
    # Mianownik + Dopełniacz + Miejscownik
    '(?:S|s)tycz(?:eń|nia|niu)',        # January
    '(?:L|l)ut(?:y|ego|ym)',            # February
    '(?:M|m)ar(?:zec|ca|cu)',           # March
    '(?:K|k)wie(?:cień|tnia|tniu)',     # April
    '(?:M|m)aj(?:a|u)?',                # May
    '(?:C|c)zerw(?:iec|ca|cu)',         # June
```

```
    '(?:L|l)ip(?:iec|ca|cu)',           # July
    '(?:S|s)ierp(?:ień|nia|niu)',       # August
    '(?:W|w)rze(?:sień|śnia|śniu)',     # September
    '(?:P|p)aździernik(?:a|u)?',        # October
    '(?:L|l)istopad(?:a|ie)?',          # November
    '(?:G|g)rud(?:zień|nia|niu)',       # December
]

day_regex = '(?:3[0-1]|[0-2]?\d)'
year_regex = '(?:2\d{3}|1?\d{1,3})'

date_regex = rf"{day_regex}\s(?:{'|'.join(months_regex)})(?:\s{year_regex}|\
```

In [6]:
```
print(time_regex)
```

```
\b(?:(?:2[0-4])|(?:1?\d)):(?:[0-5]\d)(?::[0-5]\d)?
```

In [7]:
```
print(date_regex)
```

```
(?:3[0-1]|[0-2]?\d)\s(?:(?:S|s)tycz(?:eń|nia|niu)|(?:L|l)ut(?:y|ego|ym)|(?:M
|m)ar(?:zec|ca|cu)|(?:K|k)wie(?:cień|tnia|tniu)|(?:M|m)aj(?:a|u)?|(?:C|c)zer
w(?:iec|ca|cu)|(?:L|l)ip(?:iec|ca|cu)|(?:S|s)ierp(?:ień|nia|niu)|(?:W|w)rze
(?:sień|śnia|śniu)|(?:P|p)aździernik(?:a|u)?|(?:L|l)istopad(?:a|ie)?|(?:G|g)
rud(?:zień|nia|niu))(?:\s(?:2\d{3}|1?\d{1,3})|\b)
```

In [8]:
```
example = 25906
```

In [9]:
```
corpus[example]
```

Out[9]: '„Aby uniknąć szczypców, z góry stwierdzam, że ta odpowiedź dotyczy USA; Eu
ropejczycy, Azjaci, Kanadyjczycy itp. mogą mieć zupełnie inne systemy i zas
ady. Nie musisz się martwić, jeśli spłacisz wyciąg z karty kredytowej w peł
na w dniu, w którym jest należna w odpowiednim czasie. Z drugiej strony, je
śli rutynowo utrzymujesz saldo z miesiąca na miesiąc lub bierzesz zaliczki
gotówkowe, to dokonanie dowolnej płatności, którą chcesz dokonać w tym mies
iącu, jak najszybciej pozwoli Ci zaoszczędzić więcej na finansach opłaty, n
iż kiedykolwiek mógłbyś zarobić na koncie oszczędnościowym. Ale jeśli spłac
isz w całości saldo każdego miesiąca, bardzo uważnie przeczytaj drobnym dru
kiem, kiedy płatność jest wymagalna: może to oznaczać, że płatności otrzyma
ne przed 17:00 będą być wysłane tego samego dnia, lub może powiedzieć przed
godziną 15:00 lub przed godziną 19:00 czasu wschodniego, lub w południe PST
itp. itd. Tak jak mówi JoeTaxpayer, jeśli możesz zapłacić on-line z gwarant
owanym dniem za transakcję (i robisz go przed jakimkolwiek terminem nałożon
ym przez wystawcę karty kredytowej), wszystko w porządku Twój bank pozwala
mi na wypisywanie czeków „elektronicznych” na swojej stronie internetowej,
ale czek papierowy jest wysyłany pocztą do firmy obsługującej kartę kredyto
wą. Bank twierdzi, że jeśli określę termin płatności, wyślą czek z takim wy
przedzeniem, aby firma obsługująca karty kredytowe otrzymała go w terminie,
ale czy naprawdę ufasz, że USPS dostarczy czek do południa, czy cokolwiek i
nnego? Poza tym bank wstrzyma te pieniądze w dniu wypłaty czeku. (Nie zadał
em sobie trudu, aby sprawdzić, czy trzymane pieniądze nadal przynoszą odset
ki, czy nie). W każdym razie bank zrzeka się wszelkiej odpowiedzialności za
następstwa (opłaty za opóźnienia w płatnościach, opłaty finansowe za wszyst
kie zakupy itp.), jeśli ten papierowy czek nie zostanie otrzymany na czas,
a zatem konto karty kredytowej przejdzie do „„opóźnionej płatności” " stan.
Aha, i mój bank również chce mieć miesięczną opłatę za usługę BillPay (dowo
lna liczba takich „elektronicznych” czeków dozwolona w każdym miesiącu). Us
ługa BillPay obejmuje płatności elektroniczne lokalnym sprzedawcom i przeds
iębiorstwom użyteczności publicznej, które mają konta w banku i zarejestrow
ały się, aby otrzymywać płatności drogą elektroniczną. Wszystkie moje firmy
obsługujące karty kredytowe zezwalają mi na korzystanie z ich witryn intern
etowych w celu upoważnienia ich do pobrania określonej przeze mnie płatnośc
i z mojego konta bankowego. Mogę wybrać dzień, kwotę i z którego z moich ko
nt bankowych będą pobierać pieniądze, ale muszę to robić co miesiąc. Bardzo
wygodnie pokazują kalendarz do wyboru terminu z wyraźnie zaznaczonym termin
em płatności, a jak wskazuje komentarz mhoran_psprep, płatność można zaplan
ować z dużym wyprzedzeniem przed datą, w której płatność zostanie faktyczni
e dokonana, czyli nie Nie musisz się martwić, że z powodu podróży nie będzi
esz mieć dostępu do Internetu, a tym samym nie będziesz mógł zalogować się
na stronie karty kredytowej, aby dokonać płatności w terminie. Mogę również
zarejestrować się w usłudze AutoPay, która pobiera ustaloną kwotę/minimalną
należną płatność/płatność w całości (niezależnie od tego, co wybiorę) w nal
eżnym terminie, i będzie się to odbywać miesiąc po miesiącu, bez koniecznoś
ci podejmowania dalszych działań z mojej strony. W obu przypadkach to od fi
rmy obsługującej karty zależy odebranie pieniędzy z mojego konta w określon
ym dniu, a jeśli zepsuje, nie może naliczyć opłat za opóźnienia w płatności
ach ani opłat za finansowanie nowych zakupów itp. Ponadto, w przeciwieństwi
e do mojego banku, istnieją brak opłat za tę usługę. Warto również zauważy
ć, że wielu osobom nie podoba się pomysł, aby firma obsługująca karty kredy
towe wypłacała pieniądze z ich konta bankowego, a więc ta opcja nie wszystk
im odpowiada”.'

In [10]:
```
regex.findall(time_regex, corpus[example])
```

Out[10]:
```
['17:00', '15:00', '19:00']
```

In [11]:
```
grouping_regex = "|".join(["(" + x + ")" for x in months_regex])
```

In [12]:
```
example = 29306
```

In [13]:
```
corpus[example]
```

Out[13]:
```
'Nie - w grę wchodzą dodatkowe czynniki. Należy pamiętać, że akcje wyemitow
ane przez firmę mogą ulec zmianie z różnych powodów (takich jak konwersja/u
morzenie zamiennych papierów wartościowych, nabywanie uprawnień do ogranicz
onych akcji pracowniczych, konwersja opcji pracowniczych, programy skupu ak
cji pracowniczych, lokowanie akcji, wykupy akcji, fuzje, prawa poboru itp.
.) więc zawsze warto sprawdzić ogłoszenia SEC dla firmy, jeśli chcesz uzysk
ać dokładną liczbę. Może również istnieć wiele klas akcji i uprzywilejowany
ch papierów wartościowych, które mają różne poziomy dywidend. W przypadku P
FG złożyli 10 kw. w dniu 22 kwietnia 2015 r. i zauważyli, że mają w obrocie
294 385 885 akcji zwykłych. Zauważyli również, że za trzy miesiące zakończo
ne 31 marca 2014 r. dywidendy były wypłacane zarówno akcjonariuszom zwykły
m, jak i akcjonariuszom uprzywilejowanym oraz że były akcje uprzywilejowane
serii A (3 miliony) i akcje uprzywilejowane serii B (10 milionów) oraz oświ
adczenie: W lutym W 2015 r. nasza Rada Dyrektorów zatwierdziła program wyku
pu akcji o wartości do 150,0 mln USD naszych pozostających w obrocie akcji
zwykłych. Akcje odkupione w ramach tych programów są księgowane jako akcje
własne, wykazywane według kosztu i odzwierciedlane jako zmniejszenie kapita
łu własnego. W związku z tym dokładna kwota wypłaconej dywidendy będzie zna
na dopiero w kolejnym raporcie kwartalnym, który będzie zawierał dokładną k
wotę dywidendy wypłaconej akcjonariuszom zwykłym i uprzywilejowanym za dany
kwartał.'
```

In [14]:
```
grouping_regex
```

Out[14]:
```
'((?:S|s)tycz(?:eń|nia|niu))|((?:L|l)ut(?:y|ego|ym))|((?:M|m)ar(?:zec|ca|c
u))|((?:K|k)wie(?:cień|tnia|tniu))|((?:M|m)aj(?:a|u)?)|((?:C|c)zerw(?:iec|c
a|cu))|((?:L|l)ip(?:iec|ca|cu))|((?:S|s)ierp(?:ień|nia|niu))|((?:W|w)rze(?:
sień|śnia|śniu))|((?:P|p)aździernik(?:a|u)?)|((?:L|l)istopad(?:a|ie)?)|((?:
G|g)rud(?:zień|nia|niu))'
```

In [15]:
```
regex.findall(grouping_regex, 'luty')
```

Out[15]:
```
[('', 'luty', '', '', '', '', '', '', '', '', '', '')]
```

In [16]:
```
regex.findall(grouping_regex, corpus[example])
```

Out[16]:
```
[('', '', '', '', 'maj', '', '', '', '', '', '', ''),
 ('', '', '', 'kwietnia', '', '', '', '', '', '', '', ''),
 ('', '', '', '', 'maj', '', '', '', '', '', '', ''),
 ('', '', 'marca', '', '', '', '', '', '', '', '', ''),
 ('', 'luty', '', '', '', '', '', '', '', '', '', '')]
```

In [17]:
```
print(date_regex)
```

```
(?:3[0-1]|[0-2]?\d)\s(?:(?:S|s)tycz(?:eń|nia|niu)|(?:L|l)ut(?:y|ego|ym)|(?:M
|m)ar(?:zec|ca|cu)|(?:K|k)wie(?:cień|tnia|tniu)|(?:M|m)aj(?:a|u)?|(?:C|c)zer
w(?:iec|ca|cu)|(?:L|l)ip(?:iec|ca|cu)|(?:S|s)ierp(?:ień|nia|niu)|(?:W|w)rze
(?:sień|śnia|śniu)|(?:P|p)aździernik(?:a|u)?|(?:L|l)istopad(?:a|ie)?|(?:G|g)
rud(?:zień|nia|niu))(?:\s(?:2\d{3}|1?\d{1,3})|\b)
```

In [18]:
```
regex.findall(date_regex, corpus[example])
```

Out[18]:
```
['22 kwietnia 2015', '31 marca 2014']
```

In [19]:
```
def hour_extractor(hit: str) -> int:
    """
    Format is either HH:MM:SS or HH:MM
    :param hit: Time regex match
    :return: hour as int
    """
    times = hit.split(":")
    assert len(times) == 2 or len(times) == 3, hit
    return int(times[0])
```

In [20]:
```
def month_extractor(hit: str) -> int:
    """

    :param hit: Date regex match
    :return: month as a int
    """
    grouping_regex = "|".join(["(" + x + ")" for x in months_regex])
    return [i for i, match in enumerate(regex.findall(grouping_regex, hit)[0
```

In [21]:
```
hours_str = [
    hour_extractor(match)
    for text in corpus
    for match in regex.findall(time_regex, text)
]
```

In [22]:
```
months_str = [
    month_extractor(match)
    for text in corpus
    for match in regex.findall(date_regex, text)
]
```

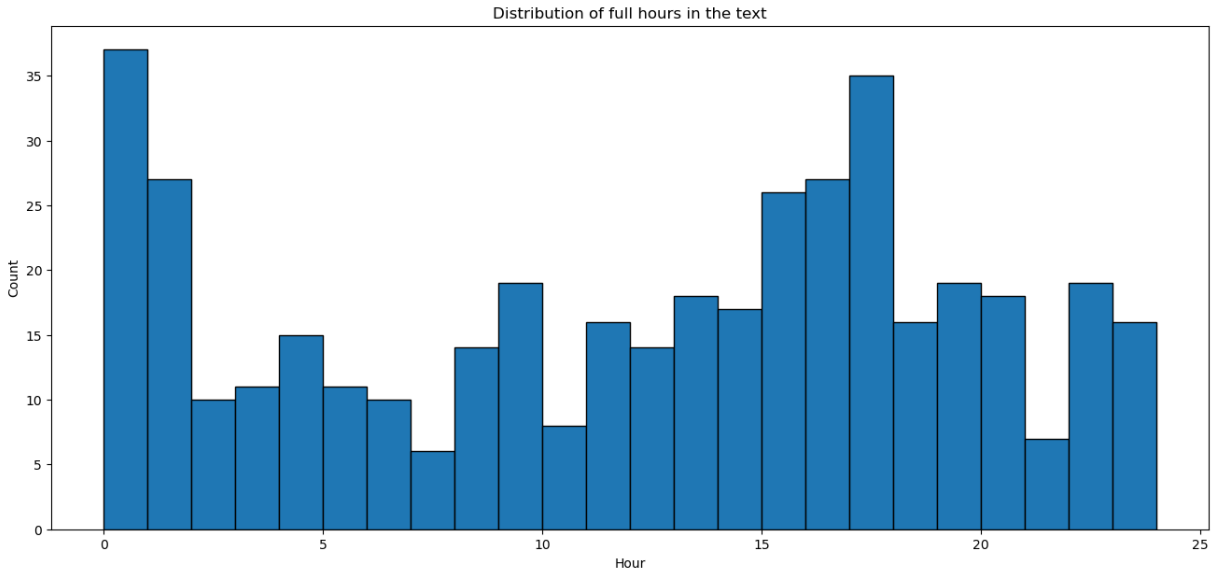
In [23]:
```
len(hours_str)
```

Out[23]:
```
416
```

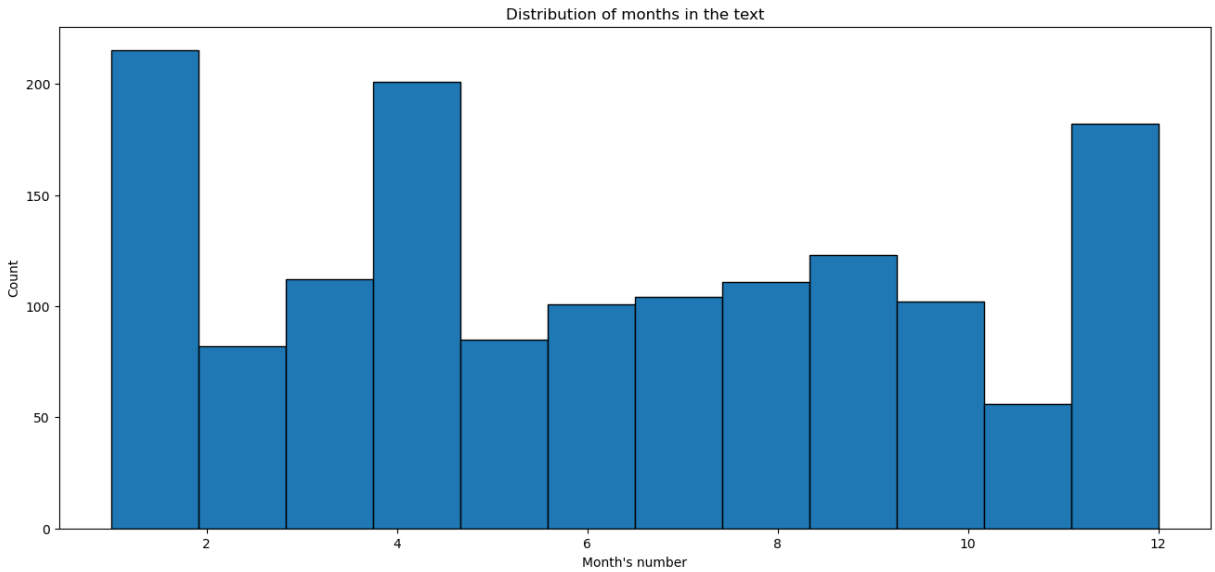
In [24]:
```
len(months_str)
```

Out[24]:
```
1474
```

In [25]:
```
plt.hist(hours_str, edgecolor='black', bins=24)
plt.title('Distribution of full hours in the text')
plt.ylabel('Count')
plt.xlabel('Hour')
plt.savefig('pictures/hours_dist.svg')
```



In [26]:
```
plt.hist(months_str, edgecolor='black', bins=12)
plt.title('Distribution of months in the text')
plt.ylabel('Count')
plt.xlabel("Month's number")
plt.savefig('pictures/months_dist.svg')
```



In [27]:
```
january_regex = r"\bstycz(?:eń|ni(?:owi|em|u|e|om|ami|ach|a|ów)?)\b"
sum(len(regex.findall(january_regex, text, regex.IGNORECASE)) for text in co
```

Out[27]: 463

In [28]:
```
number_with_january = rf"\d+\s{january_regex}"
sum(len(regex.findall(number_with_january, text, regex.IGNORECASE)) for text
```

Out[28]: 215

In [29]:
```
january_without_number = rf"(?<!\d+\s){january_regex}"
sum(len(regex.findall(january_without_number, text, regex.IGNORECASE)) for t
```

Out[29]: 248

It sums up :)

Questions

---

1. Are regular expressions good at capturing times?

- They can handle regular times quite good, without any problems. But catching times in mixed or in verbal form might be problematic (like e.g. `po 20 wieczorem`)

2. Are regular expressions good at capturing dates?

- If we want to capture dates in standard numerical format (like `YYYY-MM-DD` or similar) they can be useful, but not in case of verbal notation. Especially for languages with declension (like Polish), they can be quite problematic, huge and unreadable

3. How one can be sure that the expression has matched all and only the correct expressions of a given type?

- If we're looking for standard data types in programming languages we can cast results to expected type, and no error should arise. Additionally, we can write unit tests for specific regex (or ask chat GPT for that :))
- We can also try some sort of rubber duck debugging, or explaining our regex to another person - which I used for this excersice

In [ ]: