Techniki analizy sieci społecznych
Projekt 2: Raport końcowy

Maksim Makaranka 308826
Jakub Sprawka 315533
Kamil Sulkowski 310936

Jak wygląda sieć powiązań europosłów zrekonstruowana na podstawie ich współobecności na spotkaniach z lobbystami?

Interpretacja tematu

Celem projektu jest zrekonstruowanie sieci powiązań pomiędzy członkami Komisji Europejskiej na podstawie ich współobecności na spotkaniach z lobbystami, z dodatkowym uwzględnieniem ich wieku, przynależności partyjnej oraz ukończonego kierunku studiów. Analizując, którzy członkowie komisji uczestniczą w tych samych spotkaniach oraz jakie mają cechy, możemy zidentyfikować potencjalne grupy wpływów, koalicje czy wspólne zainteresowania wynikające z podobieństw w wieku, przynależności politycznej czy wykształceniu. Projekt polega na stworzeniu grafu społecznego, w którym wierzchołkami są członkowie komisji z przypisanymi atrybutami, a krawędziami połączenia między nimi, reprezentujące liczbę wspólnych spotkań z lobbystami.

Zbieranie i przetwarzanie danych

Poniżej znajduje się szczegółowa dokumentacja procesu zbierania danych w naszym projekcie, podzielona na podrozdziały według zastosowanych scraperów.

Scraper spotkań lobbystów z członkami Komisji Europejskiej

Plik: `lobbyist_meetings_html_scraper.py`
Klasa: `LobbyistMeetingsScraper`

Ten scraper służy do zebrania danych o spotkaniach każdego lobbysty w czasie kadencji Komisji Europejskiej 2019-2024. Wykorzystuje bibliotekę *Selenium* do automatyzacji przeglądarki oraz *BeautifulSoup4* do parsowania stron internetowych.

Proces działania:

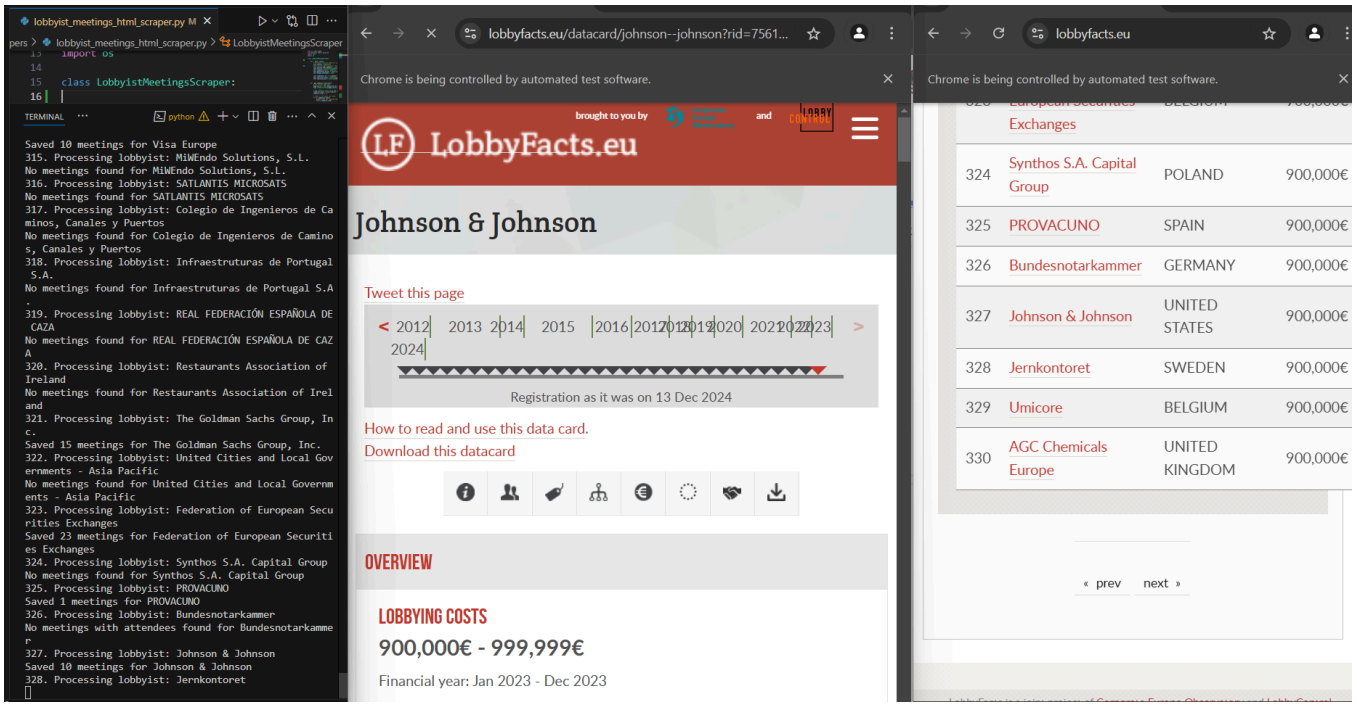
1. **Uruchomienie dwóch instancji ChromeDriver'a:**
   - **Pierwszy ChromeDriver:** Nawiguje po stronach z lobbystami na stronie https://www.lobbyfacts.eu/ i zbiera linki do poszczególnych lobbystów z każdej strony, utrzymując paginację.
   - **Drugi ChromeDriver:** Otwiera zebrane linki do profili lobbystów i pobiera dane o ich spotkaniach. Używając *BeautifulSoup4*, parsuje treść stron oraz weryfikuje daty spotkań, aby upewnić się, że mieszczą się w okresie kadencji Komisji 2019-2024.
2. **Zapisywanie danych:**
   - Dane są zapisywane do plików JSON, nazywanych zgodnie z nazwami lobbystów.
   - Struktura danych w plikach JSON jest następująca:

```
{
    "meetings": [
      ["Członek A", "Członek B"],
      ["Członek A", "Członek C", "Członek D"],
      ...
    ]
}
```

Każda lista wewnątrz "meetings" reprezentuje jedno spotkanie i zawiera nazwiska członków Komisji, którzy w nim uczestniczyli.

3. **Tworzenie listy unikatowych uczestników:**
   - Po zakończeniu procesu scrapowania generowany jest plik `__unique_attendees.json`, zawierający listę unikatowych uczestników spotkań wraz z ich stanowiskami. Będziemy się na nim opierać w kolejnych etapach projektu.

W trakcie działania programu można zaobserwować dwa otwarte okna ChromeDriver'a (Rys. 1): jedno z paginacją na stronie https://www.lobbyfacts.eu/, drugie z profilem skrapowanego lobbysty. Logi w konsoli pokazują przebieg działania scrapera, co ułatwia monitorowanie postępu i ewentualne debugowanie.



Rysunek 1: Przebieg działania scrapera LobbyistMeetingsScraper.

Scraper danych o członkach Komisji Europejskiej z pliku PDF

Plik: `commissioners_data_pdf_scraper.py`
Klasa: `CommissionersDataScraper`
Łańcuch: `commissioners_data_chain.py` (klasa `CommissionersDataChain`)

Ten scraper korzysta z pliku PDF dostępnego pod tym adresem. Wykorzystuje bibliotekę *LangChain* do przetwarzania języka naturalnego przy zastosowaniu modeli OpenAI.

Proces działania:

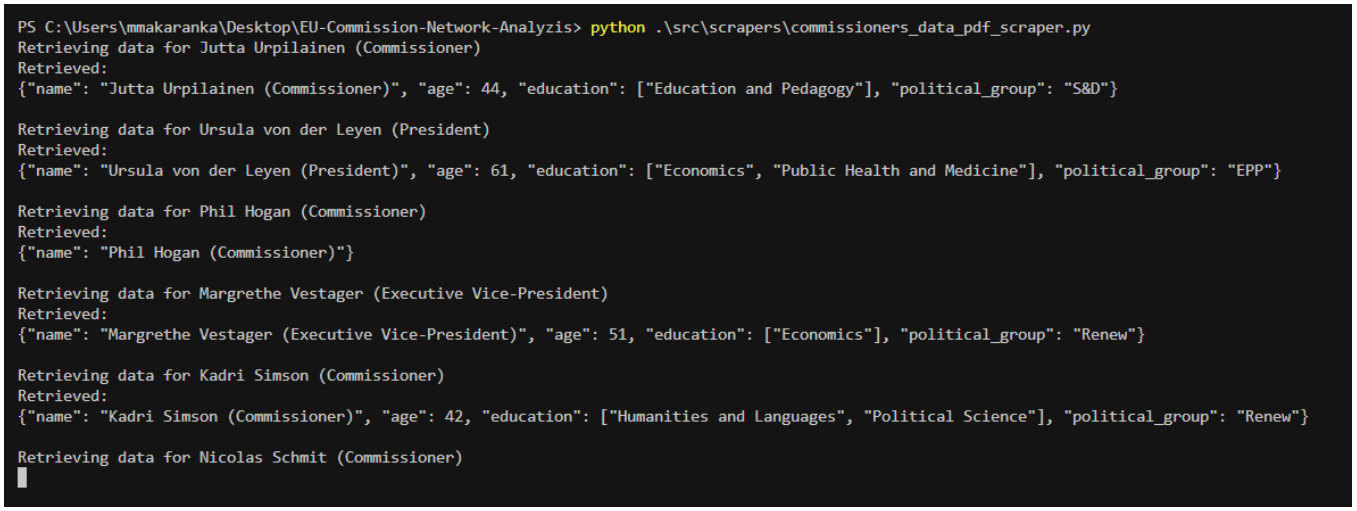
1. **Tworzenie bazy wektorowej *FAISS*:**
   - Tworzymy bazę wektorową *FAISS* zawierającą embeddingi dla 27 stron dokumentu źródłowego z opisami członków Komisji. Embeddingi są generowane za pomocą modelu OpenAI *ada v2*.
2. **Wyszukiwanie i ekstrakcja danych:**
   - Vectorstore jest wykorzystywany do odnalezienia odpowiedniej strony raportu dla danego komisarza poprzez porównanie podobieństwa tekstowego.
   - Tworzony jest prompt dla modelu *gpt-4o-mini*, który na podstawie wyszukanych danych zbiera informacje o komisarzach i zapisuje je do pliku JSON o następującej strukturze:

```
{
    "commissioners": [
      {
        "name": "Imię i Nazwisko (Stanowisko)",
        "age": 60,
        "education": [
          “Kierunek 1”,
          “Kierunek 2”,
          “...”
        ]
        "political_group": "Grupa"
      },
      ...
    ]
}
```

3. **Użyty prompt:**

```
<RETRIEVED-DATA>
{retrieved_data}
</RETRIEVED-DATA>
Based on the retrieved information above, determine the age, education, and political group data for the commissioner named {commissioner}.
Select one or more education fields from the list below:
<EDUCATION-FIELDS>
# List of manually selected education fields (see code)
</EDUCATION-FIELDS>
Return the data in the following JSON format:
{
   "age": 45,
   "education": ["Engineering", "Economics"],
   "political_group": "EPP"
}
If the retrieved data does not provide any relevant information about the specified commissioner, return empty JSON.
```

Podczas działania scrapera w konsoli wyświetlane są logi, które ilustrują proces ekstrakcji danych z pliku PDF (Rys. 2).



Rysunek 2: Logi działania scrapera CommissionersDataScraper.

Scraper dopasowujący członków gabinetów do członków Komisji

Plik: `cabinet_members_match_perplexity_html_scraper.py`
Klasa: `CabinetMembersMatchScraper`
Łańcuch: `cabinet_member_matching_chain.py` (klasa `CabinetMemberMatchingChain`)

Podczas analizy danych zauważyliśmy, że większość spotkań odbywa się z udziałem członków gabinetów komisarzy lub innych urzędników Komisji Europejskiej, takich jak Director General, Director, Head of Task Force czy Head of Service, a nie samych komisarzy. Naszym zadaniem jest wyfiltrowanie członków gabinetów oraz dopasowanie ich do odpowiednich komisarzy.

Proces działania:

1. **Filtrowanie członków gabinetów:**
   - Filtrowanie jest przeprowadzane na podstawie informacji o stanowiskach, które są umieszczone w nawiasach przy nazwiskach uczestników spotkań. Zostawiamy tych, które mają "Cabinet member".
2. **Dopasowanie członków gabinetów do komisarzy:**
   - Ze względu na brak wiarygodnego i wszechstronnego źródła danych o przynależności członków gabinetów, podjęto różne próby pozyskania tych informacji.
   - Rozważano skrapowanie danych z LinkedIn, jednak po analizie okazały się niewystarczające. Podjęto próbę wykorzystania silnika Tavily do przeszukiwania internetu, jednak w większości przypadków zwracał on komunikat o braku poszukiwanej informacji.
3. **Wykorzystanie strony perplexity.ai:**
   - Postanowiono wykorzystać stronę perplexity.ai, która korzysta z modeli LLM z możliwością przeszukiwania internetu.
   - Sprawdzono, że model udostępniany przez perplexity.ai często udziela wiarygodnych i sprawdzalnych odpowiedzi na pytania dotyczące przynależności członków gabinetów do komisarzy.
   - Podjęto próbę wykorzystania API Perplexity, jednak okazało się, że nie oferuje ono funkcji przeszukiwania internetu, a jedynie dostęp do zfinetunowanych modeli *llama-3.1-sonar* stosowanych przez Perplexity. W związku z tym zdecydowano się na bezpośrednie skrapowanie danych ze strony perplexity.ai.
4. **Skrapowanie strony perplexity.ai:**

- Za pomocą *Selenium* oraz *BeautifulSoup4* opracowano scraper, który automatyzuje interakcję ze stroną perplexity.ai.
- Program uruchamia stronę perplexity.ai i czeka, aż użytkownik się zaloguje oraz naciśnie Enter w konsoli, ponieważ zaobserwowano, że jakość wyników wyszukiwania jest lepsza po zalogowaniu.
- Iterując po członkach gabinetów, do perplexity.ai wysyłany jest prompt:

```
If you know, specify which Commissioner's Cabinet {cabinet_member} serves to during the
2019-2024 EU Commission. Return Cabinet and Commissioner full name.
```

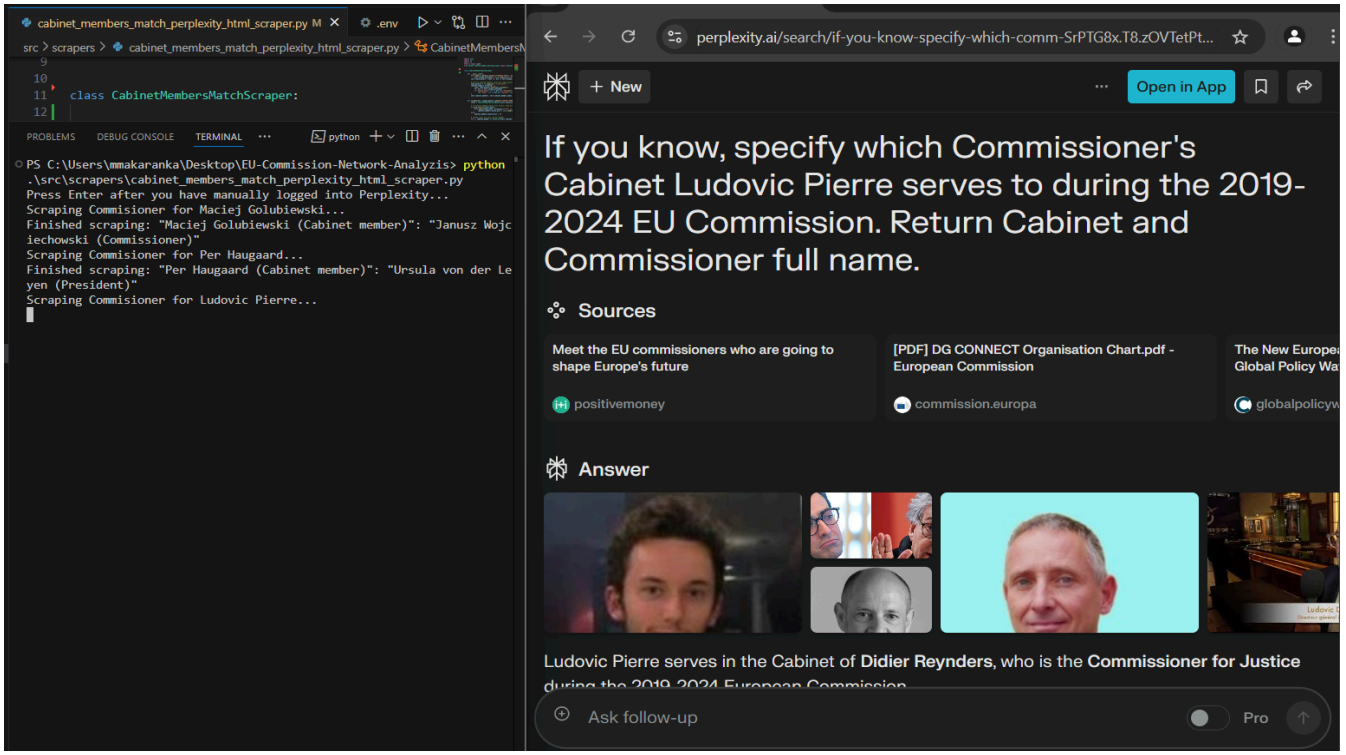
- Otrzymana odpowiedź jest następnie analizowana przy użyciu modelu *gpt-4o-mini*, z wykorzystaniem następującego promptu:

```
<RETRIEVED-DATA>
{retrieved_data}
</RETRIEVED-DATA>
Based on the retrieved information above, identify and return the record of the European
Commissioner whose cabinet includes the member {cabinet_member} from the list below. If the
retrieved data does not contain the Commissioner from the list, return UNKNOWN.
<COMMISSIONERS-LIST>
{commissioners_list}
</COMMISSIONERS-LIST>
<RESPONSE-FORMAT>
{
   "commissioner": "Imię Nazwisko (Stanowisko)"
}
</RESPONSE-FORMAT>
<UNKNOWN-FORMAT>
{
   "commissioner": "UNKNOWN"
}
</UNKNOWN-FORMAT>
```

- Wynikowy komisarz jest przypisywany do członka gabinetu w słowniku, który jest zapisywany do pliku `cabinet_members_match.json`.

5. **Rozwiązanie problemów z CAPTCHA:**
   - Strona perplexity.ai wykrywała automatyzację i wyświetlała CAPTCHA od Cloudflare.
   - Aby obejść ten problem, wykorzystano bibliotekę *undetected-chromedriver*, która pozwala na ukrycie faktu automatyzacji przed mechanizmami antybotowymi.

Podczas działania scrapera można obserwować w konsoli logi, które pokazują proces wysyłania zapytań i analizowania odpowiedzi (Rys. 3). Dodatkowo, otwarte jest okno przeglądarki z uruchomioną stroną perplexity.ai, co umożliwia monitorowanie interakcji z serwisem w czasie rzeczywistym.



Rysunek 3: Przebieg działania scrapera CabinetMembersMatchScraper.

Budowa i charakterystyka grafów

Do stworzenia grafu została wykorzystana biblioteka *networkx* języka **Python.**

Bazą do tworzenia grafu były następujące pliki zebrane przez scrapera:

- *commisioners_data_reviewed.json*
- *cabinent_members_match.json*
- Wszystkie pliki w folderze *meetings*

Założenia do budowy grafu

- Członkowie komisji bardzo rzadko pojawiają się osobiście na spotkaniach z lobbystami, dlatego obecność członka gabinetu członka komisji traktujemy jego obecność.
- Kolejnym rzadkim zjawiskiem jest pojawianie się reprezentacji dwóch różnych członków na dokładnie tym samym spotkaniu, dlatego w celu stworzenia grafu zawierającego ciekawe dane, jako jedno spotkanie traktujemy zbiór wszystkich spotkań z danym lobbystą. Pozwala to stworzyć więcej powiązań przy zachowaniu istotnych informacji (obecność członka komisji na spotkaniu z lobbystą pokazuje, że reprezentowane przez niego sprawy są istotne dla członka). Stworzenie grafu biorąc pod uwagę tylko konkretnie to samo spotkanie też może dostarczyć ważnych informacji dlatego postanowiliśmy stworzyć 3 wersje grafu:
    - **Graf 1: Waga krawędzi między dwoma członkami jest równa liczbie lobbystów na których spotkaniach oboje byli, niekoniecznie w tym samym momencie – wtedy dla sytuacji:**

LOB1: [K1, K2], [K1], [K3], [K5] ; LOB2: [K2, K3], [K1], [K4]

(gdzie notacja LOBx: [K1, K2], [K3] oznacza że lobbysta LOBx odbył dwa spotkania z członkami komisji lub ich przedstawicielami, pierwsze z członkami K1 i K2 a drugie z członkiem K3)

krawędź między K1 i K2 będzie miała wagę 2, bo obaj uczestniczyli w obu spotkaniach, a między K1 i K5 wagę 1 bo obaj uczestniczyli tylko w spotkaniach z LOB1

- **Graf 2: waga krawędzi podobnie jak w Grafie 1 zależy od obecności na spotkaniach z tym samym lobbystą ale również od liczby obecności – wtedy dla:**

  LOB1: [K1, K2], [K1], [K3], [K5] ; LOB2: [K2, K3], [K1], [K4]

  waga krawędzi z K1 do K2 będzie równa 2,5 (1,5 poprzez spotkania z LOB1, bo K1 był na 2 spotkaniach, a K2 na jednym [(2 + 1) / 2] i 1 ze spotkań z LOB 2 [(1 + 1) / 2]

- **Graf 3: waga krawędzi zależy tylko od wspólnych spotkań, tylko w tym samym czasie – wtedy dla przykładu:**

  LOB1: [K1, K2], [K1], [K3], [K5] ; LOB2: [K2, K3], [K1], [K4]

  krawędzie powstaną tylko dla K1 i K2 oraz K2 i K3, ponieważ tylko oni uczestniczyli we wspólnych spotkaniach, a wagi krawędzi zależą od liczby wspólnych spotkań

- Nie wszyscy uczestnicy spotkania należą do gabinetu badanych członków komisji, dlatego próba przyporządkowania im członka komisji z naszej listy była niemożliwa dla modelu LLM, co prowadziło do zjawiska halucynacji i przydzielania im błędnych członków komisji lub osoby niezwiązanej. By zapobiec zniszczeniu danych, osoby takie są odfiltrowane, wykorzystując ich stanowisko lub w skrajnych przypadkach ręcznie, poprzez ich imię i nazwisko (Joseph Vella i Fiona Knab-Lunny).
- Podczas spotkania obecność więcej niż jednego przedstawiciela danego gabinetu jest traktowana jakby był tylko 1. Przyjmujemy takie założenie ponieważ liczba przedstawicieli nie wydaje się nam by miała zbieżność z zainteresowaniem tematem.

Proces tworzenia grafu

- Stworzenie pustego grafu i dodanie do niego wierzchołków z członkami komisji i ich atrybutami (wiek, wykształcenie i przynależność partyjna)

```
def create_graph_with_members() -> nx.Graph:
    graph = nx.Graph()
    with open(COMMISIONERS_DATA_DIR, "r", encoding="utf8") as f:
        data = json.load(f)["commissioners"]

    for commissioner in data:
        name = commissioner["name"]
        del commissioner["name"]
```

```
            graph.add_node(name, **commissioner)
    return graph
```

- Dla każdego pliku opisującego spotkania dla lobbysty wyciągamy wszystkich unikalnych uczestników spotkania, po czym konwertujemy ich na członków komisji, których reprezentują (członek komisji oczywiście reprezentuje sam siebie)

```
def extract_all_unique_members(file):
    if "__unique_attendees.json" in file:
        return None
    with open(file, "r", encoding="utf8") as f:
        data = json.load(f)["meetings"]
    unique_attendees = set()
    for meeting in data:
        for attendee in meeting:
            unique_attendees.add(attendee)

    result = set()
    members_list = all_members_list()
    match = members_match()
    for attendee in unique_attendees:
        if attendee in members_list:
            result.add(attendee)
        else:
            try:
                result.add(match[attendee])
            except KeyError:
                if attendee.split("(")[1].replace(")", "") in UNREALTED_TITLES:
                    pass
                      elif "Joseph Vella" in attendee or "Fiona Knab-Lunny" in
attendee:
                    pass
                else:
                    print(f"missing {attendee}")
    return result
```

Lista *UNRELATED_TITLES* zawiera tytuły osób niezrzeszonych z żadnym członkiem, więc są pomijane w celu zachowania sensowności danych. Podobnie dzieje się w przypadku dwóch wcześniej wymienionych członków.
Kod ten przeznaczony jest do tworzenia Grafu nr 1. Dla grafów 2 i 3 jest on niemal identyczny z delikatnymi różnicami, więc przedstawienie metod mija się z celem.

- Ze wszystkich unikalnych członków tworzone są wszystkie kombinacje dwuelementowe, i jeśli nie istnieje między nimi krawędź, jest ona tworzona z wagą 1, a jeśli istnieje, to jej waga jest

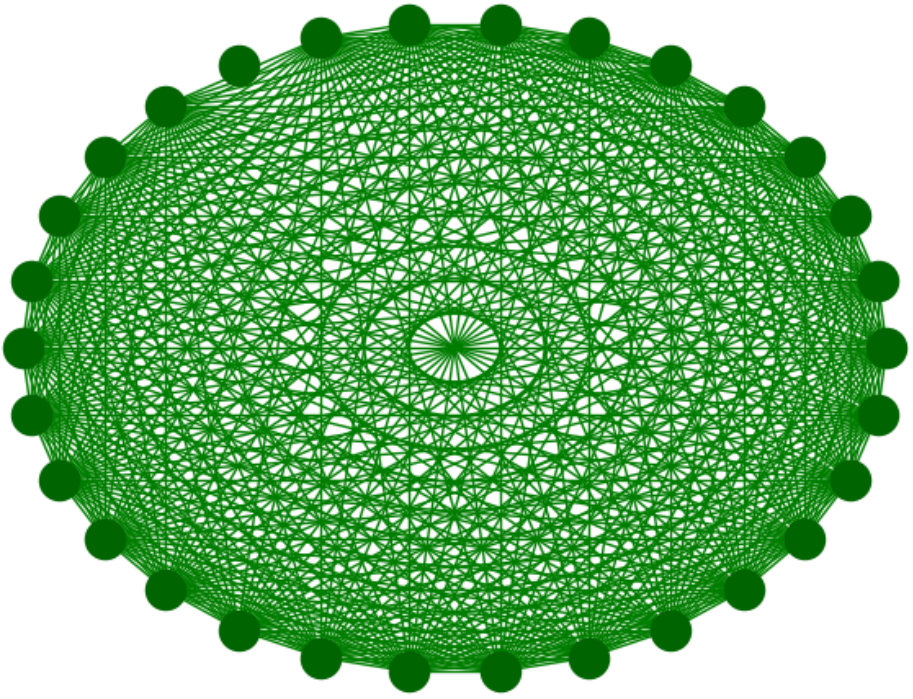
zwiększana o 1.

```
def create_edges_for_meeting(file, graph: nx.Graph):
    if "__unique_attendees.json" in file:
        return None
    attendees = extract_all_unique_members(file)
    for combination in combinations(attendees, 2):
        a1, a2 = combination
        if graph.has_edge(a1, a2):
            graph[a1][a2]["weight"] += 1
        else:
            graph.add_edge(a1, a2, weight=1)
```

- Ostatnim krokiem jest usunięcie niepotrzebnych wierzchołków, czyli w tym wypadku krawędzi *UNKNOWN* (ostatni krok: usunięcie uczestników bez przypisania).

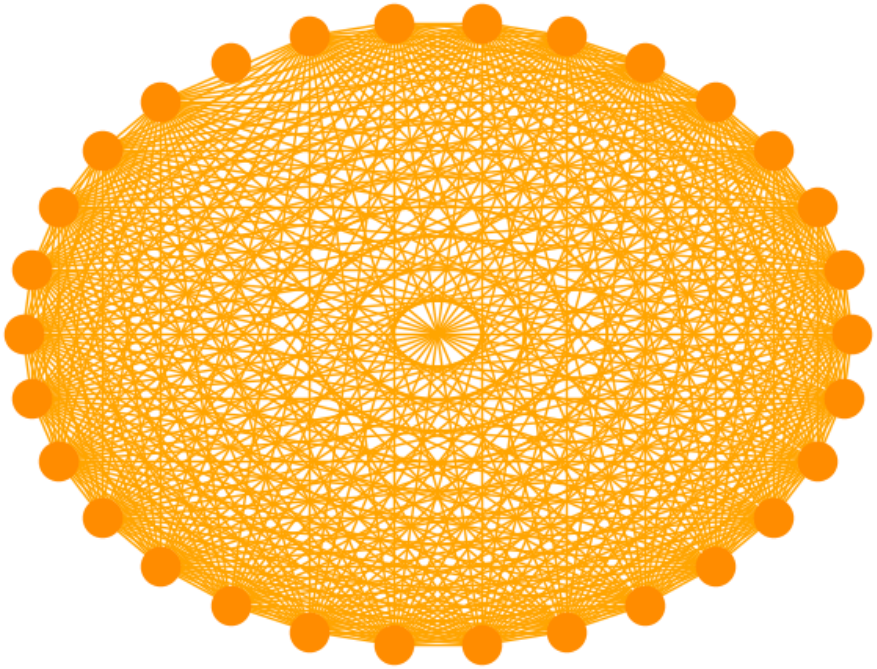
Rezultatem są **3 grafy**:

- **Graf nr 1** (Rys. 4): 30 wierzchołków i 406 krawędzi (graf pełny, pomijając jeden wierzchołek, który nie ma żadnych połączeń),
- **Graf nr 2** (Rys. 5): Jedyną różnicą z grafem 1 jest fakt, że wagi jego krawędzi są większe,

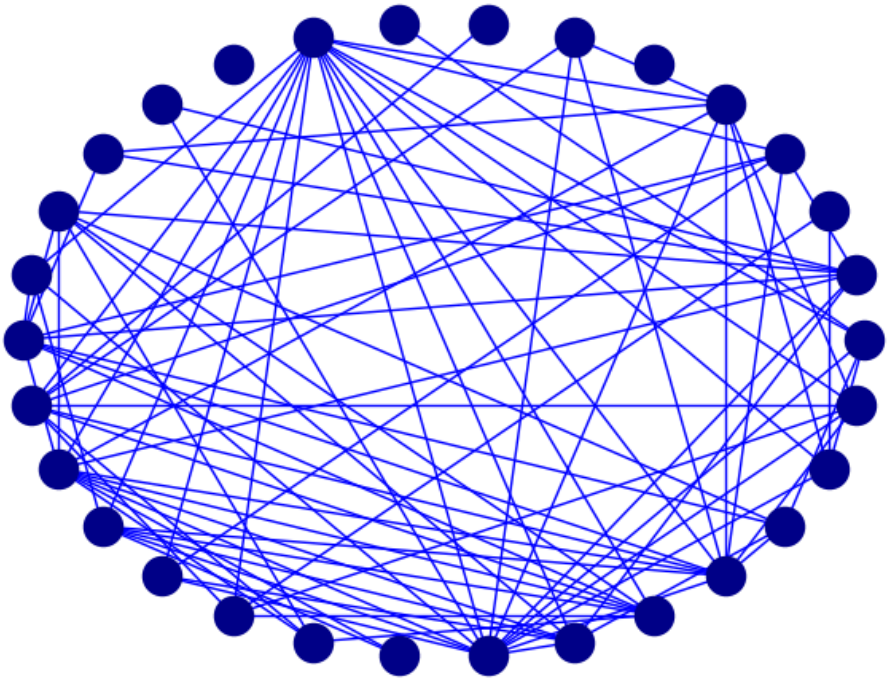


Rysunek 4: Wizualizacja kołowa grafu nr 1.

- **Graf nr 3** (Rys. 6): Graf ten też ma 30 wierzchołków, ale tylko 89 krawędzi. Tym razem jednak (pomijając członka bez żadnych spotkań) graf nie jest spójny, Helena Dalli (Commissioner) nie miała żadnych wspólnych spotkań. Wszyscy pozostali członkowie tworzą graf spójny.



Rysunek 5: Wizualizacja kołowa grafu nr 2.

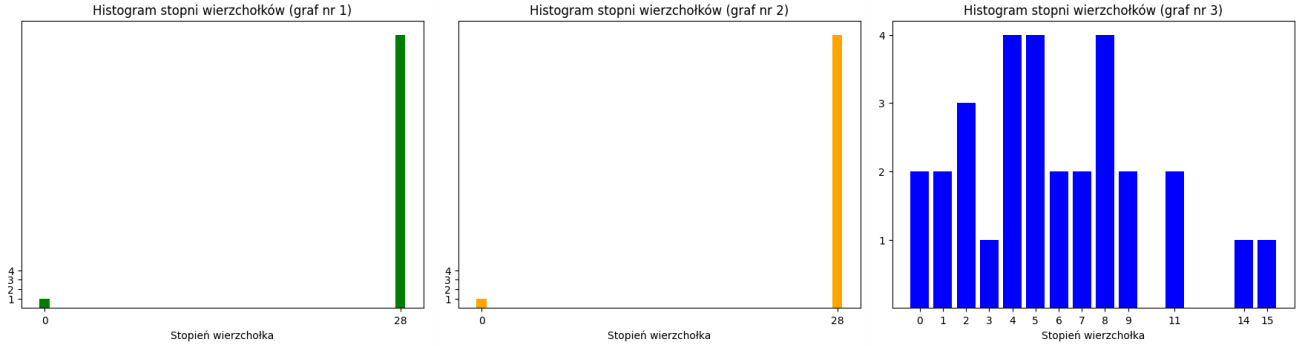


Rysunek 6: Wizualizacja kołowa grafu nr 3.

Analiza grafów

Po stworzeniu docelowych grafów przystąpiliśmy do ich analizy zgodnie z punktami ustalonymi w pierwszym etapie projektu. Dzięki uzyskaniu trzech modeli byliśmy w stanie dokonać pełnej analizy ze względu na różne charakterystyki grafów. Analizy dokonaliśmy przy użycia notatnika w pliku *src/graphs/experiments.ipynb* oraz głównie biblioteki *networkx* przy wsparciu modułu *scipy* oraz *numpy*.

Analiza stopni wierzchołków



Rysunek 7: Histogramy stopni wierzchołków dla trzech grafów.

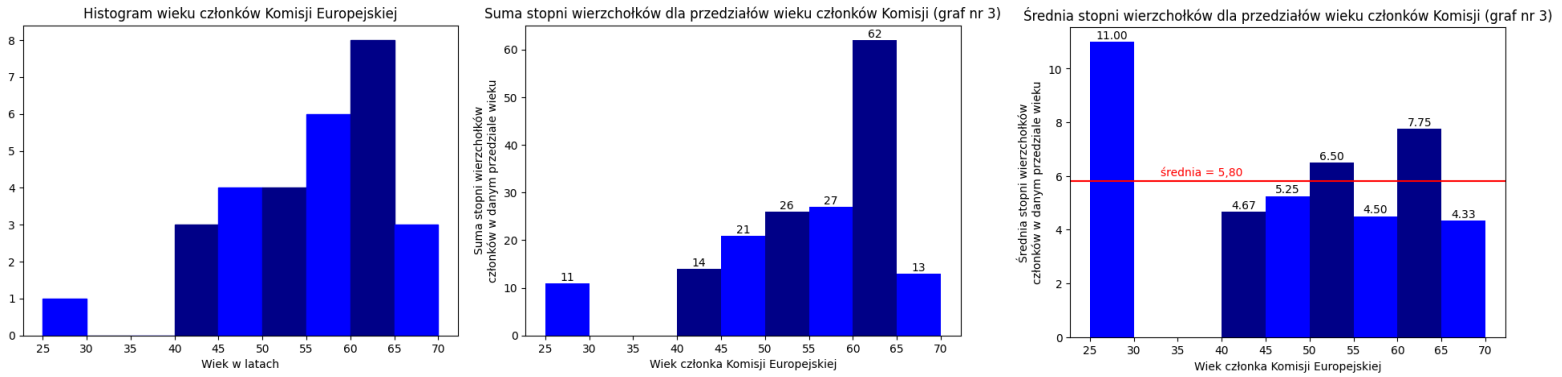
Grafy nr 1 i 2, po usunięciu jednego wierzchołka nie posiadającego żadnego połączenia z resztą grafu – Josep Borrell (Vice-President) – są grafami pełnymi, zatem analiza z wykorzystaniem stopni ich wierzchołków bez uwzględnienia wag krawędzi mija się z celem (Rys. 7).

Ze względu na sposób budowy grafu nr 3, możemy zidentyfikować członków Komisji najbardziej powiązanych z pozostałymi (ze względu na stopień wierzchołków) oraz spróbować zbadać korelację wieku członków komisji ze stopniem ich wierzchołków w grafie.

Czterech najbardziej powiązanych z resztą członków Komisji to:

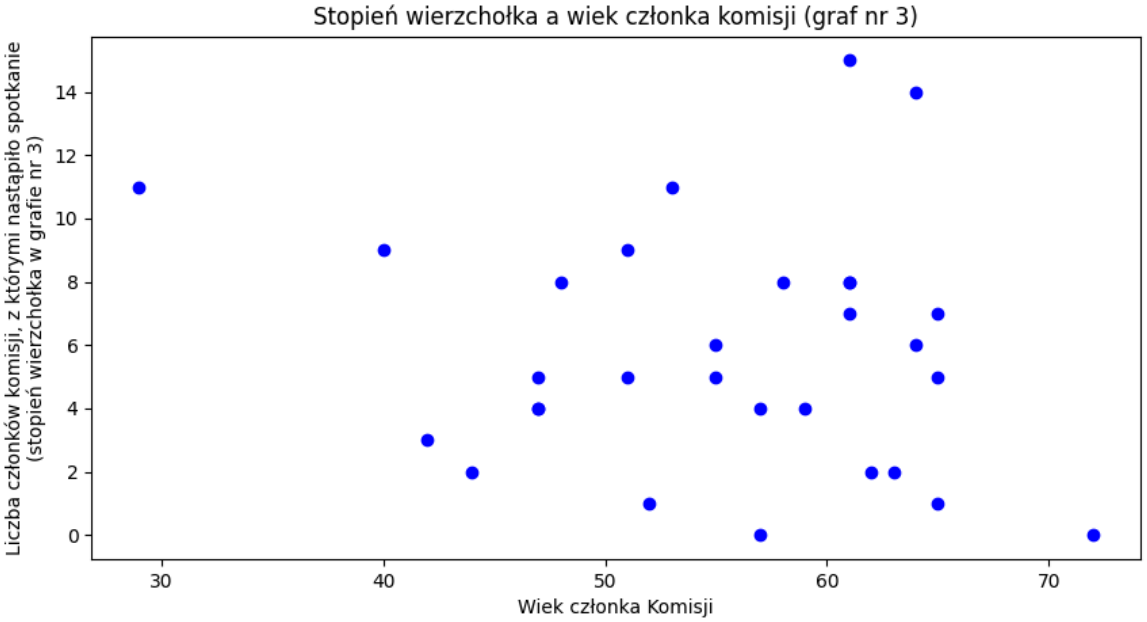
- Johannes Hahn (Commissioner): 15,
- Thierry Breton (Commissioner): 14,
- Maroš Šefčovič (Vice-President): 11,
- Virginijus Sinkevičius (Commissioner): 11.

Widać również, że mamy dwóch członków, którzy nie są powiązani z resztą grafu. Do Josepa Borrella doszła Helena Dalli (Commissioner).



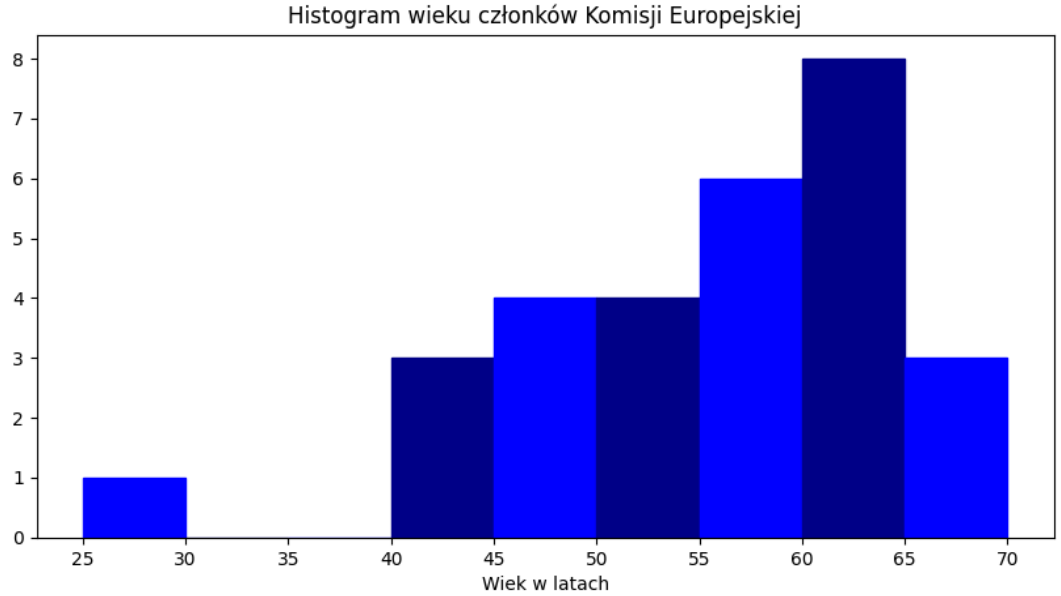
Rysunek 8: Histogram wieku członków oraz wykresy sumy i średniej stopni wierzchołków od przedziałów wieku członków Komisji dla grafu nr 3.

W celu zbadania korelacji, wykreśliliśmy wykres stopnia wierzchołka od wieku członka Komisji (Rys. 9). Następnie wyliczyliśmy współczynnik korelacji dystansowej, którego wartość równa 0,14 wskazuje na znikomą zależność obu zmiennych. Również analiza samego wykresu nie wskazuje na żaden związek wieku członka z liczbą powiązań z pozostałymi.



Rysunek 9: Wykres rozkładu punktów (stopień wierzchołka, wiek członka Komisji) dla grafu nr 3.

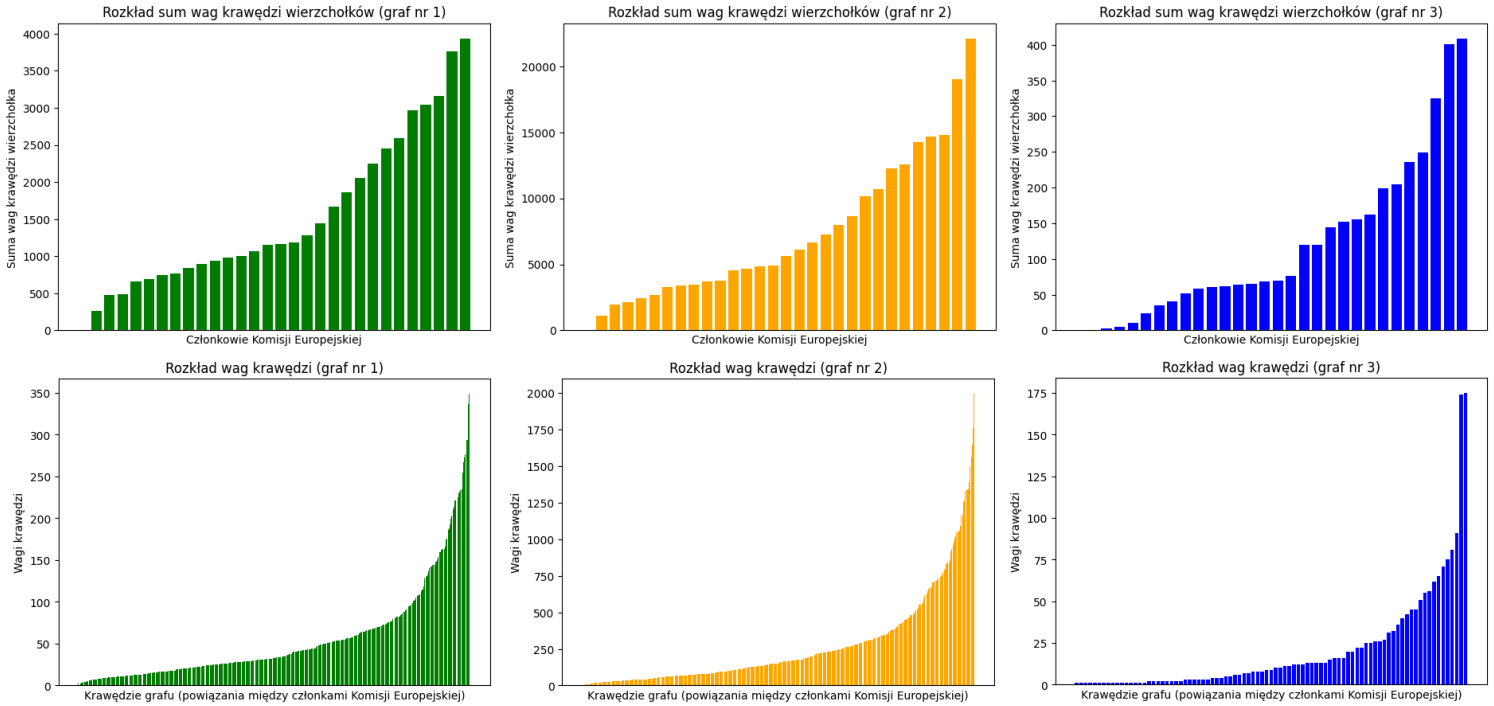
Postanowiliśmy podzielić członków ze względu na przedziały wiekowe, aby móc porównać histogram ich wieku z sumą stopni ich wierzchołków w grafie nr 3 (Rys. 8, 10). Duże podobieństwo obu grafów oraz wysoki współczynnik korelacji Perasona między wartościami w koszykach histogramu wieku oraz sumy stopni wierzchołków (wartość 0,94) sugeruje, iż nie ma głębszej zależności między obiema zmiennymi. Z wykresu średnich stopni wierzchołków w przedziałach wiekowych rzuca się w oczy odstająca wartość dla pierwszego przedziału wieku (25-30 lat), w którym znajduje się jeden członek Komisji.



Rysunek 10: Histogram wieku członków Komisji Europejskiej

Analiza stopni wierzchołków z uwzględnieniem wag krawędzi

W kolejny kroku przystąpiliśmy do analizy wag krawędzi grafów. Zbadaliśmy rozkłady wag pojedynczych krawędzi, jak i sum wag krawędzi dla poszczególnych wierzchołków, wykreślając odpowiednie wykresy (Rys. 11) oraz wyznaczając podstawowe statystyki (Tab. 1).



Rysunek 11: Rozkłady sumy wag krawędzi wierzchołków oraz wag krawędzi dla trzech grafów.

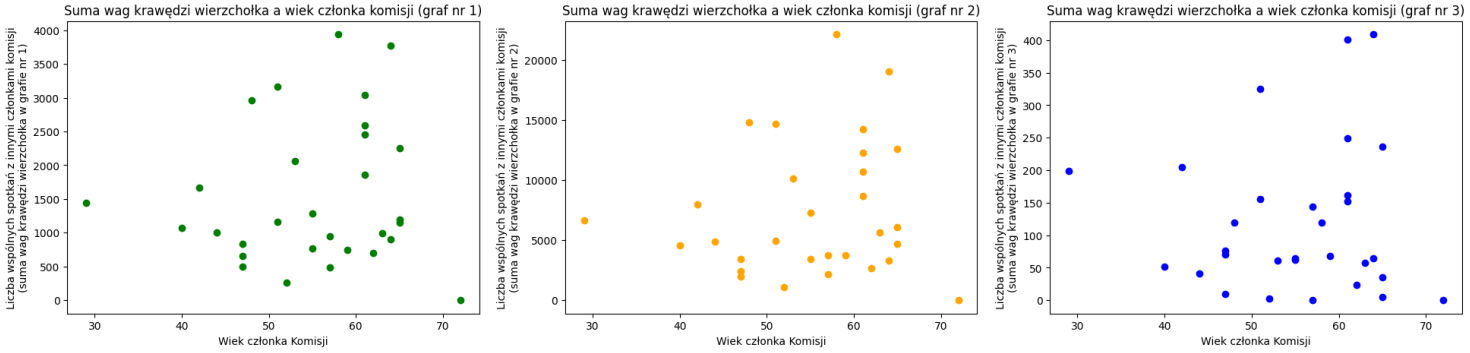
Zauważyliśmy, że rozkład wag krawędzi jest typem rozkładu z długim ogonem oraz że w każdym grafie można wyodrębnić małą grupę krawędzi o największych wagach, co może nam sugerować, że istnieją pary członków, którzy szczególnie często biorą udział we wspólnych spotkaniach (lub ich przedstawiciele). Daje to nadzieję, że uda się wykryć pewne związki między takimi wierzchołkami grafu.

	Graf nr 1		Graf nr 2		Graf nr 3	
	Sumy wag krawędzi dla wierzchołków	Wagi krawędzi	Sumy wag krawędzi dla wierzchołków	Wagi krawędzi	Sumy wag krawędzi dla wierzchołków	Wagi krawędzi
Średnia	1527,67	56,44	7325,03	270,63	119,07	20,53
Mediana	1157	33	5275,75	151,75	69	9
Odchylenie standardowe	1031,81	61,00	5425,83	321,43	111,33	31,40
Wartość minimalna	0	2	0	2,5	0	1
Wartość maksymalna	3937	349	22104	1995,5	409	175

Tabela 1: Statystyki rozkładów wag krawędzi

Podobnie jak w poprzednim punkcie, postanowiliśmy zbadać związek sumy wag krawędzi dla wierzchołków, czyli liczby wspólnych spotkań z innymi członkami Komisji, z wiekiem danego członka.

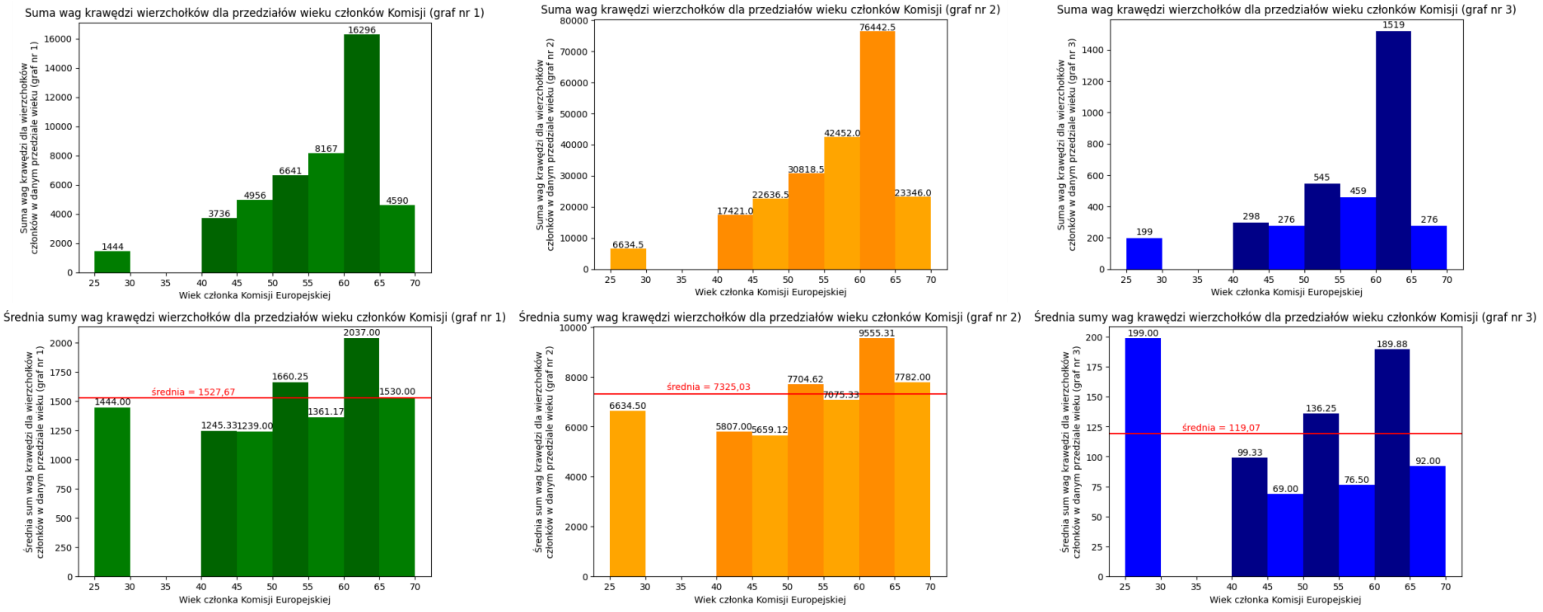
Zaczęliśmy od analizy rozkładu punktów (suma wag krawędzi wierzchołka, wiek członka), lecz podobnie jak w poprzednim punkcie nie udało nam się zaobserwować mocnych zależności między zmiennymi, co potwierdziły również wartości współczynnika korelacji dystansowej, odpowiednio dla grafów 1, 2 i 3: -0,084, -0,104, -0,014 (Rys. 12).



Rysunek 12: Wykresy rozkładu punktów (suma wag krawędzi węzła, wiek członka) dla trzech grafów.

Analogicznie pogrupowaliśmy również dane członków według ich wieku do koszyków o zakresie 5 lat ([25; 30), [30; 35), …, [60; 65), [65; 70]) i obliczyliśmy średnią oraz zsumowaliśmy wagi krawędzi ich wierzchołków w grafie (czyli liczbę spotkań z lobbystami, na których spotkali się z innymi członkami Komisji Europejskiej) (Rys. 13).

Na pierwszy rzut oka można zauważyć, że wykresy sum wag krawędzi są do siebie bardzo zbliżone jeśli chodzi o rozkłady danych. Wydaje się, że po uwzględnieniu wag krawędzi dalej nie udało się zaobserwować związku między powiązaniami członków Komisji a ich wiekiem, gdyż wykresy sumy wag przypominają rozkład wieku członków, podobnie jak to było z czystymi stopniami wierzchołków. Aby upewnić się co do naszych spostrzeżeń, wyznaczyliśmy również współczynniki korelacji Pearsona, których wartości odpowiednio dla grafów 1, 2 i 3: 0,96, 0,97 i 0,88, utwierdziły nas w przekonaniu, że ciężko doszukiwać się tutaj związku.



Rysunek 13: Wykresy sumy i średnich wag krawędzi węzłów dla przedziałów wieku dla trzech grafów.

Podobny wniosek można wysnuć z analizy wykresu średnich sum wag krawędzi wierzchołków (czyli średnich liczb wspólnych spotkań dla członków Komisji w danym przedziale wieku, Rys. 13). Najmocniej wybijają się wartości dla przedziałów wieku 60-65 oraz dla grafu nr 3, dla przedziału wieku 25-30.

W kolejnym kroku przeszliśmy do sprawdzenia krawędzi o największych wagach, czyli do par członków, którzy najczęściej spotykają się wspólnie z lobbystami. Do analizy postanowiliśmy wykorzystać graf nr 1. W Tabeli 2. przedstawiono 10 par o największej liczbie wspólnych spotkań. Rzucają się w oczy powtarzające się nazwiska członków.

Para członków Komisji	**Liczba spotkań, w których brali wspólnie udział**
Margrethe Vestager, Thierry Breton	349
Frans Timmermans, Thierry Breton	349
Frans Timmermans, Johannes Hahn	336
Johannes Hahn, Thierry Breton	294
Frans Timmermans, Margrethe Vestager	294
Thierry Breton, Ursula von der Leyen	276
Frans Timmermans, Valdis Dombrovskis	273
Thierry Breton, Valdis Dombrovskis	267
Nicolas Schmit, Thierry Breton	255

Tabela 2. Zestawienie 10 par członków komisji o największej liczbie wspólnych spotkań z lobbystami.

Postanowiliśmy zbudować podgrafy na podstawie n krawędzi o największych wagach (Rys. 14). Wstępna analiza ich wizualizacji potwierdziła nasze spostrzeżenia o powtarzających się nazwiskach wśród najpopularniejszych par członków. Postanowiliśmy zestawić charakterystyki podgrafów i m.in. znaleźć największe kliki oraz wyznaczyć centralności ich wierzchołków, korzystając z metryki centralności stopni wierzchołków, pominąwszy wagi krawędzi, gdyż w analizie bierzemy pod uwagę podgrafy zbudowane z krawędzi o n największych wagach (Tab. 3).

n	Liczba wierzchołków	Wierzchołki o największych stopniach	Największa klika	Wierzchołki o max. centralnościach
10	8	Thierry Breton: 6 Frans Timmermans: 4 Margrethe Vestager: 3	3: Thierry Breton, Frans Timmermans, Margrethe Vestager 3: Thierry Breton, Frans Timmermans, Valdis Dombrovskis 3: Thierry Breton, Frans Timmermans, Johannes Hahn	Thierry Breton: 0,86 Frans Timmermans: 0,57 Margrethe Vestager: 0,43

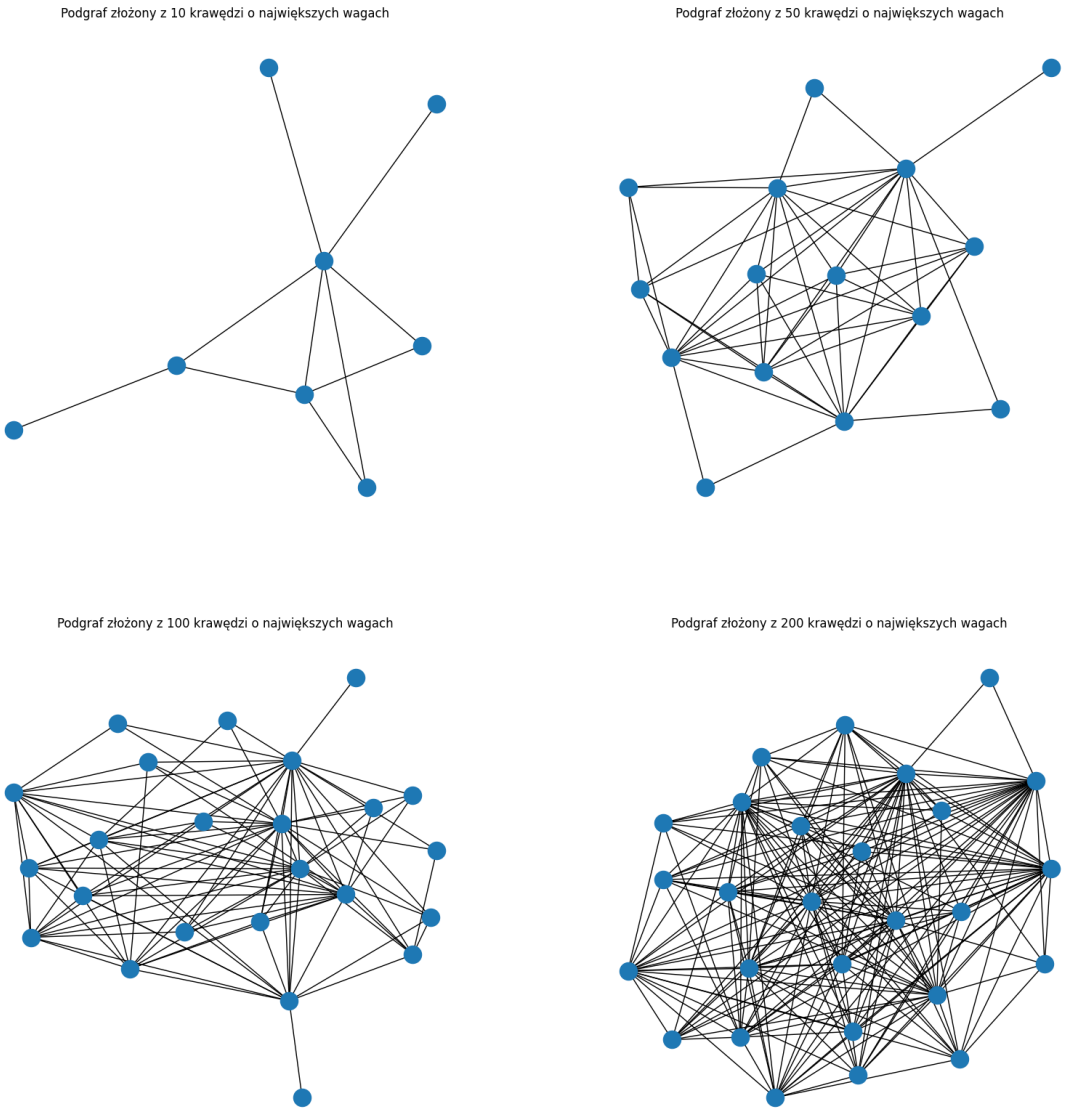
50	15	Frans Timmermans: 13 Johannes Hahn: 11 Margrethe Vestager: 11 Thierry Breton: 11	8: Frans Timmermans, Johannes Hahn, Margrethe Vestager, Valdis Dombrovskis, Thierry Breton, Ursula von der Leyen, Didier Reynders, Mairead McGuinness	Frans Timmermans: 0,93 Johannes Hahn: 0,79 Margrethe Vestager: 0,79 Thierry Breton: 0,79
100	24	Frans Timmermans: 21 Thierry Breton: 21 Johannes Hahn: 17 Margrethe Vestager:16	10: Frans Timmermans, Thierry Breton, Johannes Hahn, Margrethe Vestager, Valdis Dombrovskis, Ursula von der Leyen, Didier Reynders, Maroš Šefčovič, Mairead McGuinness, Nicolas Schmit	Frans Timmermans: 0,91 Thierry Breton: 0,91 Johannes Hahn: 0,73
200	27	Frans Timmermans: 26 Johannes Hahn: 26 Thierry Breton: 25 Margrethe Vestager: 24 Ursula von der Leyen: 24 Valdis Dombrovskis: 24	(5 klik po 12 wierzchołków)	Frans Timmermans: 1,0 Johannes Hahn: 1,0 Thierry Breton: 0,96 'Margrethe Vestager: 0,92 Ursula von der Leyen: 0,92 Valdis Dombrovskis: 0,92

Tabela 3. Zestawienie podgrafów zbudowanych z n krawędzi o największych wagach.

Możemy zaobserwować powtarzające się nazwiska we wszystkich podgrafach oraz fakt, iż udało się znaleźć największe kliki, których rozmiary rosną proporcjonalnie do liczby wierzchołków w podgrafie. Członków Komisji, którzy się w nich znajdują można uznać za grupy członków najczęściej spotykających się ze sobą na spotkaniach (lub ich reprezentantów).

Na koniec analizy związanej z wagami krawędzi sprawdziliśmy jeszcze, czy członkowie tej samej partii częściej uczestniczą w spotkaniach razem. Do tego celu obliczyliśmy średnie oraz mediany wag krawędzi dla par z tej samej partii i z różnych partii (Tab. 4).

Dla dwóch pierwszych grafów statystyki wag krawędzi są bardzo zbliżone, lecz dla grafu trzeciego, który bierze pod uwagę jedynie spotkania, w których udział wzięli dwaj członkowie (lub ich reprezentanci) jednocześnie, można zauważyć, że członkowie z tych samych partii o wiele częściej biorą udział we wspólnych spotkaniach od członków z różnych partii.



Rysunek 14: Podgrafy zbudowane na podstawie 10, 50, 100 i 200 krawędzi o największych wagach.

	Graf nr 1		Graf nr 2		Graf nr 3	
	Krawędzie par z tej samej partii	Krawędzie par z różnych partii	Krawędzie par z tej samej partii	Krawędzie par z różnych partii	Krawędzie par z tej samej partii	Krawędzie par z różnych partii
Liczba krawędzi	129	277	129	277	31	56
Średnia wag	55,95	56,67	265,67	272,94	32,10	14,13
Mediana wag	34	33	150	158	20	5

Tabela 4. Zestawienie statystyk wag krawędzi dla krawędzi par z tej samej partii oraz z różnych partii dla trzech grafów.

Analiza społeczności

Do tego celu skorzystaliśmy z implementacji metody Louvain. Wyznaczyliśmy społeczności dla wszystkich grafów i sprawdziliśmy przynależności partyjne oraz kierunki studiów ich członków:

Graf nr 1:

- 18 członków, partie: EPP: 9, S&D: 5, Renew: 4
- 11 członków, parite: S&D: 3, EPP: 5, Greens/EFA: 1, Renew: 1, ECR: 1

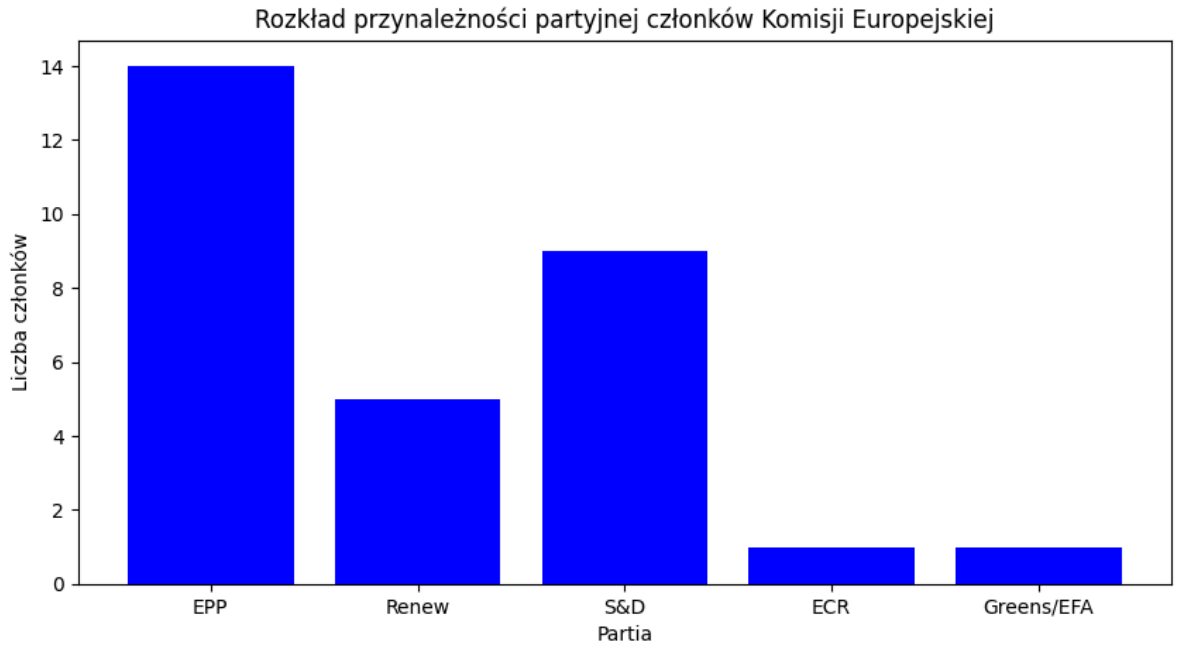
Graf nr 2:

- 10 członków, partie: S&D: 3, EPP: 4, Greens/EFA: 1, Renew: 1, ECR: 1,
- 10 członków, partie: EPP: 6, Renew: 3, S&D: 1
- 9 członków, partie: EPP: 4, S&D: 4, Renew: 1

Graf nr 3:

- 10 członków, partie: EPP: 7, ECR: 1, Renew: 1, S&D: 1
- 6 członków, partie: Renew: 1, S&D: 2, EPP: 3
- 6 członków, partie: Renew: 3, EPP: 1, S&D: 2
- 6 członków, partie: S&D: 2, EPP: 3, Greens/EFA: 1

Jak widać, w większości przypadków można zauważyć dominację jednej lub dwóch partii w utworzonych społecznościach, co może sugerować, że mają one związek z podziałami partyjnymi. Gdy przyjrzymy się rozkładowi partii dla Komisji Europejskiej (Rys. 15), można zauważyć, że dominującymi partiami są te najbardziej liczne wśród jej członków, więc związek ten nie jest naszym zdaniem dobrze udowodniony. Analiza kierunków studiów ukończonych przez członków w społecznościach nie wykazała wyraźnych trendów.



Rysunek 15: Rozkład przynależności partyjnej członków Komisji Europejskiej.

Obliczenie współczynnika klasteryzacji

Wartości wyznaczonego współczynnika klasteryzacji dla trzech grafów to odpowiednio: 0,13, 0,10 oraz 0,03. Nie zauważymy zatem tendencji do formowania się klastrów w grafach. Dodatkowo sprawdziliśmy również wartość współczynnika dla grafu nr 3, ale bez uwzględnienia wag krawędzi i choć jego wartość jest nieco wyższa (0,37), to nadal zbyt niska, aby spróbować wysnuć jakąś tezę związku tendencji do klasteryzacji danych z np. atrybutami węzłów.

Centralność wierzchołków

Aby móc zbadać centralność wierzchołków w grafach ważonych, postanowiliśmy wykorzystać współczynnik “betweenness centrality”, który obliczany jest na podstawie liczby najkrótszych ścieżek między każdymi dwoma węzłami, które przechodzą przez ten węzeł. Oto wyniki:

> **Graf nr 1** – jeden węzeł (Janez Lenarčič) z miarą centralności 0,80 oraz pozostałe węzły z wartościami współczynnika nie przekraczającymi 0,05
>
> **Graf nr 2:** – ten sam węzeł (Janez Lenarčič) z wysoką miarą centralności 0,78, jeden węzeł z niską, ale wyższą miarą od pozostałych (Iliana Ivanova – 0,19) oraz pozostałe węzły ze współczynnikiem nie przekraczającym wartości 0,07
>
> **Graf nr 3:** – trzy węzły z największą miarą centralności: Maroš Šefčovič (0,52), Virginijus Sinkevičius (0,41), Thierry Breton (0,38), pozostałe węzły z miarami poniżej 0,15

Jedynie miary centralności uzyskane dla grafu nr 3 pozwalają na zbadanie ich zależności od atrybutów węzłów o największych miarach. Niestety okazało się, że trzej członkowie o zdecydowanie największej wartości współczynnika centralności pochodzą z różnych partii (odpowiednio: S&D, Greens/EFA i EPP), a ich kierunki studiów również nie przedstawiają żadnej zależności ([Economics, International Relations, Law], , [Economics, Social Sciences], [Engineering]).

Wnioski

- Nie udało się zaobserwować korelacji między powiązaniami dotyczącymi spotkań z lobbystami wśród członków Komisji a ich wiekiem, czy skończonymi studiami. Analiza społeczności wskazała, że możemy przypuszczać, że takie podgrupy członków Komisji mogą mieć jakiś związek z podziałami partyjnymi.
- Sposób konstrukcji grafu znacząco wpływa na ujawnienie korelacji wynikających z przynależności partyjnej w sieci kontaktów. W trzecim grafie, opartym na spotkaniach, w których członkowie uczestniczyli jednocześnie, wyraźnie widać, że członkowie tej samej partii częściej biorą udział we wspólnych spotkaniach, co skutkuje wyższymi wagami krawędzi między nimi. To sugeruje silną zależność wspólnego uczestnictwa w spotkaniach od przynależności partyjnej. Natomiast w pierwszym i drugim grafie, gdzie krawędzie opierają się na spotkaniach z tymi samymi lobbystami, taka zależność nie jest obserwowana. Ostatecznie, analiza pokazuje, że uwzględnienie jednoczesnego uczestnictwa w spotkaniach w konstrukcji grafu pozwala lepiej zidentyfikować wpływ przynależności partyjnej na sieć kontaktów.
- Udało się znaleźć grupy (kliki) członków Komisji, które najchętniej biorą udział w spotkaniach z tymi samymi lobbystami (bądź ich reprezentanci).

- Wyliczane metryki oraz analizowane wykresy dla grafu nr 2 nie różniły się w znacznym stopniu od tych, dla grafu nr 1. Zatem uwzględnienie częstości występowania członków Komisji u konkretnych lobbystów (lub ich przedstawicieli), przy użyciu naszej heurystyki do budowy grafu, nie pozwoliło ujawnić dodatkowych tendencji i zależności między danymi w grafie.

Bibliografia

1. Dane o lobbystach i spotkaniach - https://www.lobbyfacts.eu/
2. Perplexity AI - perplexity.ai
3. Dane o członkach Komisji Europejskiej 2019-2024 - link
4. Dokumentacja LangChain - https://python.langchain.com/docs/
5. Dokumentacja networkx - https://networkx.org/documentation/stable/index.html