CSC 582 Spring 2023 Lab 2 Report

Authors:
Marcelo Jimenez, mjimen45@calpoly.edu
Wesley Kwok, wjkwok@calpoly.edu

Dataset Description:

Encyklopedia Solidarności (https://encysol.pl/es/encyklopedia/biogramy#page) is a digitally native internet archive of biographical sketches (*biogramy*) for members of the anti-Communist opposition in Poland from 1976-1990.

We extracted metadata as well as the descriptions of both the organizations and people using BeautifulSoup. The metadata is exceptionally useful, since it typically provides the birthdate, city of birth, and college that the person went to for the bios, as well as a short description for the orgs. When extracting the information, we also made sure to remove any extraneous styling tags, like <bold>. Some of the bios had a special birthdate and city section, but oftentimes they did not. Thus, we decided to extract the birth date, city of birth, and college through applying named entity recognition on the short description, which is formulaically structured. For example, if we came across 'ur' (which is roughly 'born' in Polish) and SpaCy tagged the next phrase as a date, we could confidently say that the date tagged is the individual's birthdate. We did the same thing for the college and city of birth.

After parsing the HTML, we wrote the parsed information to a JSON and CSV. Afterwards, we use Pandas and NX to read the information into a dataframe and perform statistical analysis and visualization.

Most Common Entities:

The first insight we tried to get was: what were the most common entities in the encyclopedia and why. To accomplish this we first created a "sents" file that merged all the biography and organization information that the html files provided had. Then we decided to use spacy to help us with Named Entity Recognition. We decided to use the Polish language model for this task. The four types of entities that we were interested in exploring were: person name, place name, organization name, and geographical name. For all the data in our "sents" file we extract the entities and if the label falls into any of the categories we tally the counts in the dictionaries. After sorting the dictionaries by highest count we export them to csv to look at the results. We filtered out some entities like "A.". The following are the 10 most common entities for the categories listed:

personName:

Karol Józef Wojtyła, better known as Pope John Paul II, was motivated by a belief that Catholicism opposed Communism's suppression of religious, economic and political freedoms. He saw Christianity as an inseparable part of Poland's rich cultural history, and sought to re-establish a society where Poles could freely embrace their national and religious identity.

Jana Pawła II	165
Jerzego Popiełuszki	153
Lecha Wałęsy	152
G.	122
Adama Mickiewicza	111
Jerzy Popiełuszko	109
Kazimierz Jancarz	107
Marii Curie-Skłodowskiej	103
Tadeusz Isakowicz-Zaleski	102
Maksymiliana	101

placeName:

Gdańsk is a port city on the Baltic coast of Poland. The word Solidarność is synonymous with the city of Gdańsk. Although Poland's first free labor union was born out of the 1980 Lenin Shipyard strikes in Gdańsk, Solidarność would bloom into a nationwide social movement.

Warszawie	1964
Gdańsku	1102
Krakowie	958
Lublinie	935
Wrocławiu	882
Szczecinie	822
Ośr	779
Poznaniu	637
Rzeszowie	573
Polski	546

orgName:

The Komitet Komitet Obywatelski (Citizens' Electoral Committee) was an initially semi-legal political organization of the democratic opposition in Communist Poland. Formed on 18 December 1988 in Warsaw, it evolved into a nationwide movement attracting a vast majority of supporters seeking political change in the country.

„S”	5500
Wydz	2864
Komitet Komitet Obywatelski	2528
Komisja Zakładowa	1437
„Solidarność”	1333
„Solidarność”	865
SB	847
Komisja Krajowa NR związek zawodowy	843
Komitetu Założycielskiego	725
SOS	673

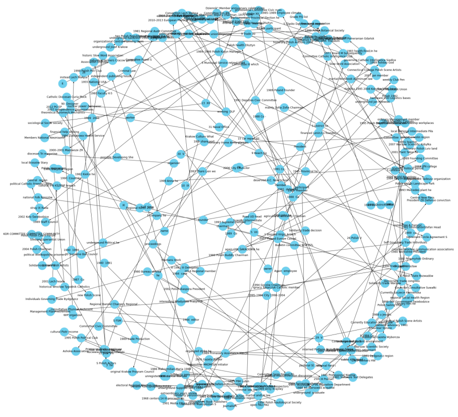
geogName:

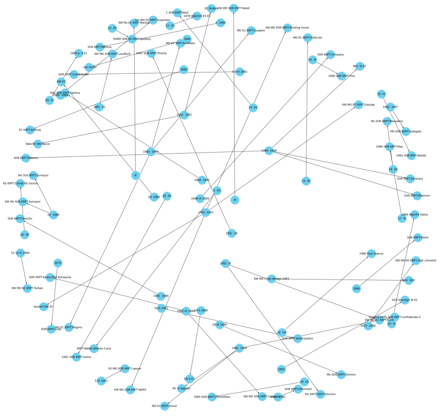
(The Cross of Freedom and Solidarity) was established on 5 August 2010, to honor members of the democratic opposition in Poland who between the years 1956 and 1989 were killed, seriously wounded or injured, arrested, imprisoned, lost jobs or were expelled from school or university for at least 6 months as a result of their activities for the benefit of a free and democratic Poland.

ul.	482
Wydz	431
kościele św.	331
parafii św.	304
Krzyżem Wolności	284
Oficerskim Orderu Odrodzenia	279
Jasną Górę	218
Krzyżem Kawalerskim Orderu Odrodzenia	117
Komandorskim Orderu Odrodzenia	116
Rodzinom	76

Relationship Extraction

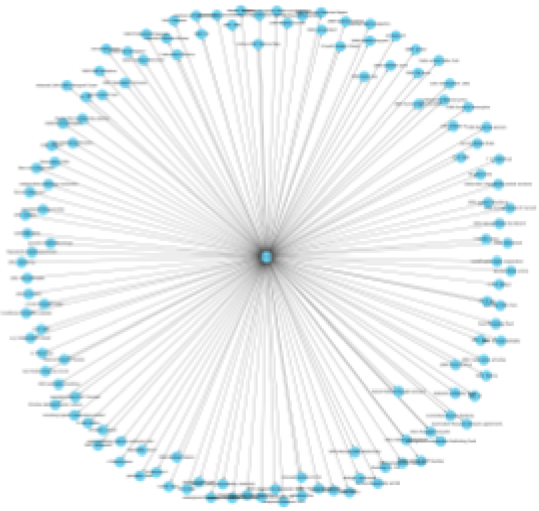
We then performed relationship extraction on the “sents” file. Our initial approach was to extract relationships from the “sents” directly. This did not work because our function to get entity pairs would struggle with parsing entities in Polish. We then tried to recreate the “sents” file but instead of keeping each sentence in Polish, we would translate them to english first, and then use coreference resolution so the named entity recognition would return the original entity and not return a pronoun. This was a very computationally expensive task given the magnitude of the dataset, so our final approach was to use only the biogramy information translated to english using the “googletrans” library. Once we perform the translation, we then extract entities and relations using the grammar of the sentences. The subject and the object will be the entity pairs, and the main verb as the relation edge. We then created a dataframe with 3 columns: source (entity1), edge (relation), target (entity2). From the data frame we export our graph’s source, edge, target counts sorted by most frequents as csv for further insight. As for edges, the most common labels we got were “graduated from”, “from”, “in”, but later we can observe some labels such as: “member of” (left graph) or “Developed by” (right graph) that show us the linkage of entities:



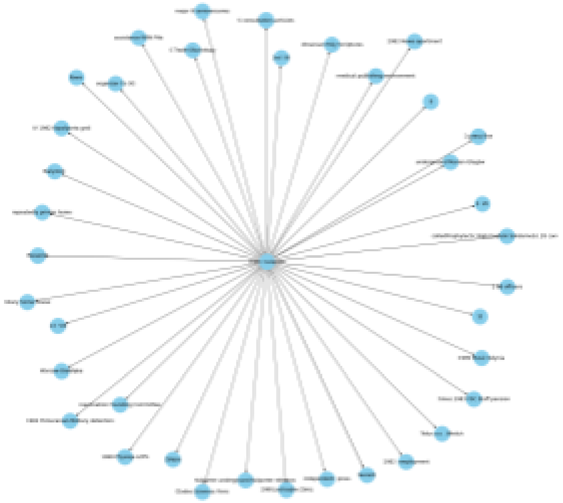


Using the “graduate of” edges, we were able to see that some common target node names included: Gdańsk Technology, Warsaw Technology, Technical School, and Bielsko Biała. Some other frequent edge labels are: “Isolation in”, “interrogated”, “detained”, “arrested” and other terms referring to the fight that the opposition was having against the government.

The most common source node was “13 XII”, which after a web search we got redirected to “Martial Law in Poland” wikipedia site. This was the date when . The government of the Polish People's Republic implemented the law to counter political opposition, in particular the Solidarity movement and thousands of opposition activists were imprisoned without trial.

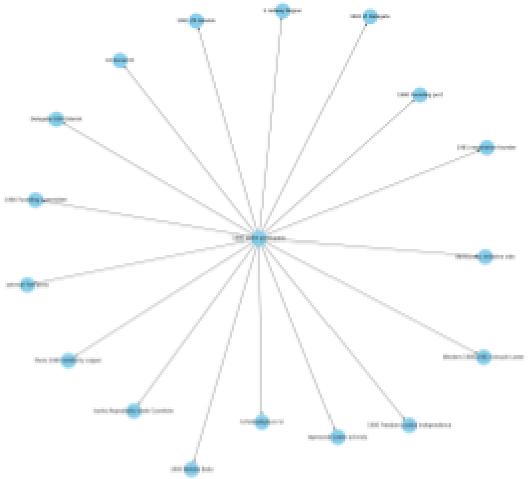


Martial Law in Poland connections

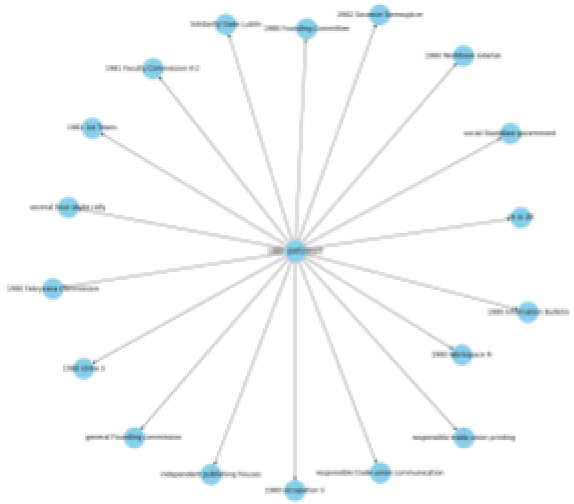


1981 Isolation Connections

Another frequent source node was “1981 Isolation” which a web search again refers us to an article “Poland imposes martial law ‘to avert anarchy’ – archive, 1981”. The majority of the target nodes are dates or places such as “Warsaw Białołęka” which is a district in the capital. Another type of source node that we found very frequent were nodes linking to the 1980 strike. Searching this event on the web, we found that it was the birth of solidarity in Poland. On 22 September 1980, Solidarity, the independent Polish trade union, was formally founded when 36 regional unions united under the name Solidarność.

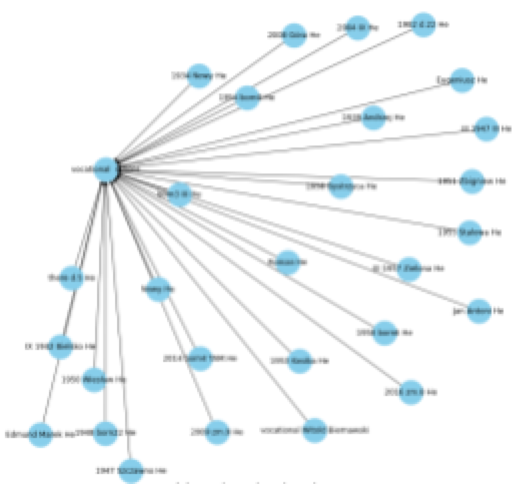


1980 strike participant

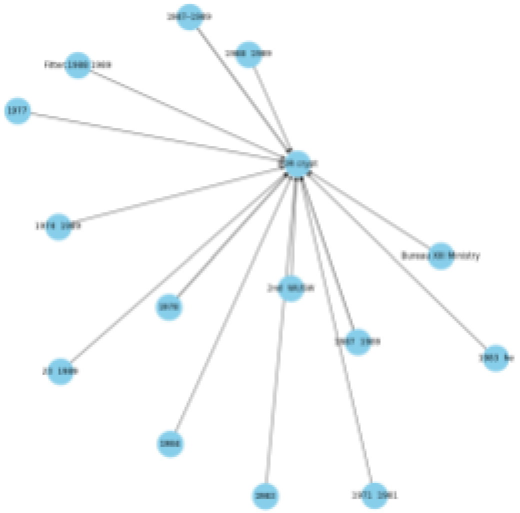


1980 Participant

As for target nodes, we found that the node with most incoming edges had the label “Polonia Restituta”. This is a Polish state order conferred on both military and civilians as well as on foreigners for outstanding achievements. Another target node with high frequency was “vocational school”. The nodes that point to this target are people like Mentzel, Zbigniew (Graduate of Warsaw University). Our participation involved in the Solidarity movement have had some sort of higher level education. The node Freedom had also many people directed to it as well as “The Cross of Valor” which is a Polish military decoration, or phrases such as “Semper fidelis”, a Latin phrase that means "always loyal". Another target node that we would like to highlight is the “Sor Crypt” or St. Leonard's Crypt. St. Leonard's Crypt under the Wawel Cathedral in Kraków, Poland, is a Romanesque crypt founded in the 11th century. The node had incoming edges from many other date nodes, however after a web search we did not find a connection to the Solidarity movement in Poland.



Vocational school



SOR Crypt

Insights regarding Colleges and Cities of Birth:

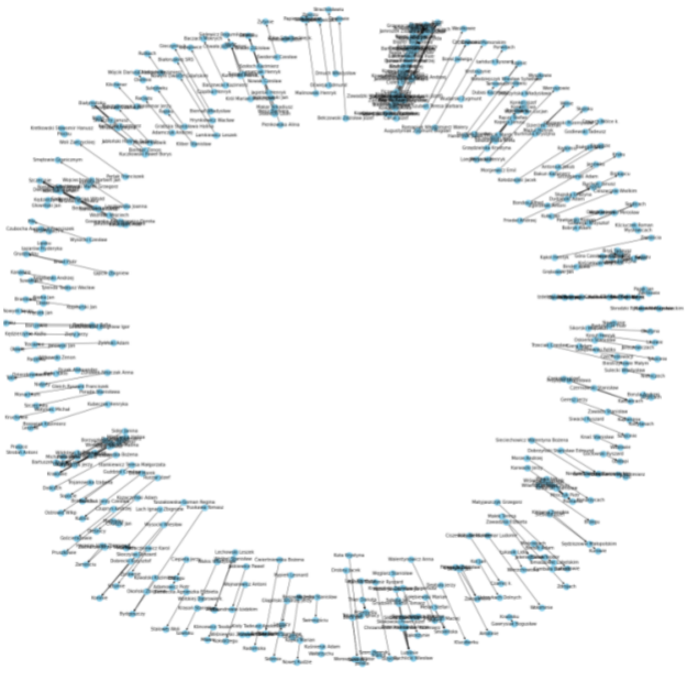


Figure A: Graph of non-college educated individuals connected to their birthplaces.

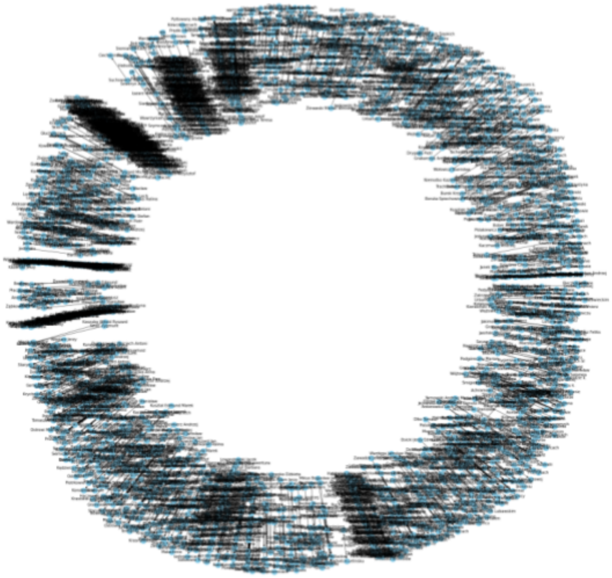


Figure B: Graph of college-educated members connected to their birthplace

To create Figure A and Figure B, we converted the parsed json info into a pandas dataframe. Afterwards, we selected the name and birthplace of each member as the source and target, filtering by non-college and college-educated members respectively for A and B. Unfortunately, the sheer amount of data points (especially on Figure B) distorts the names of many of the cities and individuals, but there are still some general insights we have gleaned from comparing these two figures.

There are many more members with some sort of post high school education.
In terms of sheer data points, comparing Figure B with Figure A shows that there are many more college-educated individuals in the Solidarity Movement in the Encyklopedia Solidarności. This fact may contribute to the Solidarity movement's organization and success.

Warsaw and Gdansk were both important cities in the movement; many of the members listed in the Encyklopedia Solidarności were born there
In addition to creating figures A and B, we also counted and sorted the birthplaces of the members by the number of members born in each city. We found that the most members were born in Warsaw and then Gdansk. Warsaw is a very large city in Poland, and a google search reveals that Gdansk was the headquarters of the movement.

A good number of college-educated members went to vocational schools instead of formal colleges
When sorting the colleges by the number of people that graduated from them, we noticed that there were many different vocational schools listed. Thus, we combined counts from all the vocational schools into a single value and re-sorted the colleges. After doing the correction, the vocational schools were number two on the list, right behind Uniwersytetu Warszawskiego - the University of Warsaw. It makes sense that many of the members were vocationally educated, since Solidarity was a trade union.

The members of Solidarity were mainly established professionals in their middle ages.
By taking all of the birth dates and averaging them out, we discovered that 1946 is the average birth year of the members. Considering that Solidarity was founded in 1980, this means that the average age of the members at founding was 34 years. Since we also know that most of these members had some sort of post-college education, we can surmise that these members were likely established professionals in their crafts, and they probably felt their livelihoods threatened by the Communist movements in Poland.

The Solidarity movement was formed of subgroups, which seem relatively organized - each with a founding committee. Oftentimes, very active members were part of several subgroups.
When parsing the data, we used SpaCy's NLP pipeline and kept a list of every organization tagged in each individual's long description. Afterwards, we combined all of the lists and sorted it by number of occurrences. We found that many active individuals participated in several of these subgroups (moving from one to another with the passage of time). One of the most common organization keywords was 'Komitetu Założycielskiego,' which roughly translates to the Founding Committee. Searching this word up in the Encyclopedia brings back many results, and in context, there seems to be a Komitetu Założycielskiego for each subgroup. Some of the most common organizations listed seem to include 'AŚ', 'KS'. Mentions of 'SOS krypt' and 'KE krypt' were also high on our list of organizational words, but we are not sure what these krypts are. Translated to English, it seems to just be 'crypt,' but that definition does not make sense in that context.