

Wydział Informatyki, Elektroniki i Telekomunikacji

Katedra Informatyki



**Analiza danych historycznych o bitwach i
zawartych pokojach w oparciu o angielską
wikipedię**

Dokumentacja

Piotr Szmigielski, Marcin Lis

Wykonano w ramach przedmiotu Zaawansowane Techniki Integracji Systemów

2015

Spis treści

1. Sformułowanie zadania projektowego	3
2. Harmonogram prac	4
3. Etapy działania	5
4. Baza danych	7
5. Szczegółowy opis tabel	8
6. Mapowanie danych źródłowych	11
7. Przetwarzanie pobranych danych	13
8. Modyfikacja modelu	14
9. Sposób przetworzenia danych	15
10. Wizualizacja i analiza danych	18
a. Podstawowe wykresy	18
b. Analizy sieci	41
11. Podsumowanie	48
Lista źródeł	48

1. Sformułowanie zadania projektowego

Celem projektu jest pozyskanie i analiza danych historycznych dotyczących bitew, wojen i pokojów na świecie w latach 1301-1800. Podany okres historyczny został wybrany ze względu na mnogość konfliktów zbrojnych pomiędzy różnymi państwami, a jednocześnie dosyć pełną wiedzę historyczną, w przeciwieństwie do okresów wcześniejszych.

Źródłem danych wykorzystywanych do analiz jest angielska Wikipedia. Dokonaliśmy tego wyboru w oparciu o takie jej zalety jak łatwość pozyskania dużych zbiorów danych, a także posiadanie w dużej mierze ustrukturyzowanych informacji, które dodatkowo zachowują relacje między poszczególnymi fragmentami danych (np. bitwa jako element wojny). Po krótkim zapoznaniu z innymi potencjalnymi źródłami danych podjęliśmy decyzję, że nie będziemy z nich korzystać, ponieważ dane są niepełne i trudne do zautomatyzowanego pobierania.

W bazie danych zapisywane są podstawowe informacje o bitwach (data, strony konfliktu, głównodowodzący armii, rozmiar sił, straty, rezultat walki). Każda bitwa jest przyporządkowana do wojny w ramach której się odbyła, z kolei do każdej wojny przypisany jest pokój (lub kilka pokoi), który ją zakończył. Pobieramy również podstawowe informacje o wojnach (czas trwania, strony konfliktu).

Od strony technicznej korzystamy z następujących technologii:

- język programowania: C#
- baza danych: relacyjna baza MS SQL
- pozyskiwanie danych: biblioteka HTML Agility Pack (parsowanie)
- persystencja: NHibernate
- analiza sieci: Gephi

Wszystkie wyżej wymienione technologie i narzędzia zostały wybrane ze względu na nasze doświadczenie z nimi, łatwość konfiguracji (baza MS SQL, NHibernate) oraz dobre dopasowanie do rozwiązywanego problemu (HTML Agility Pack, Gephi).

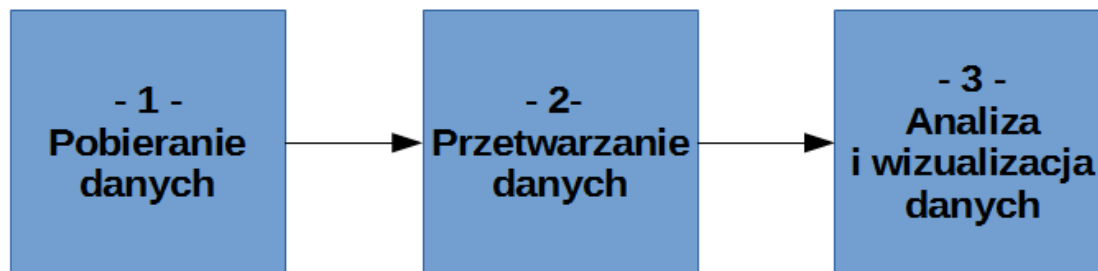
Docelowo będziemy chcieli przetworzyć zebrane dane w celu wyświetlenia ciekawych statystyk związanych z wojnami, oraz analizę relacji między poszczególnymi stronami konfliktów czy dowodzącymi bitwami.

2. Harmonogram prac

KAMIEŃ MIŁOWY	ZADANIA DO WYKONANIA	DATA ZAKOŃCZENIA
1	Przygotowanie schematu bazy danych i postawienie (relacyjna baza danych MS SQL).	27.04.2015
	Przygotowanie mapowań NHibernate i stworzenie podstawowego modelu służącego do reprezentacji danych (C#).	
	Stworzenie <u>crawlera</u> do pobierania informacji o <u>bitwach</u> i wojnach (C#).	
	<u>Pobranie</u> danych dotyczących bitew.	
2	Stworzenie <u>crawlera</u> do pobierania informacji o pokojach (C#).	18.05.2015
	<u>Pobranie</u> danych dotyczących pokoiów.	
	<u>Przetworzenie</u> wszystkich danych i uzupełnienie braków w modelu.	
3	Analiza podstawowych statystyk na podstawie zapytań SQL.	08.06.2015
	Analiza sieci i próba znalezienia w nich ciekawych zależności.	

Poszczególne etapy realizacji nie były podzielone pomiędzy autorów projektu, lecz były wykonywane wspólnie.

3. Etapy działania



POBIERANIE DANYCH

W pierwszym etapie pobrane zostaną surowe (nieprzetworzone w żaden sposób) dane pochodzące ze stron bitew, wojen i pokojów na angielskiej wikipedii. Za pobieranie odpowiedzialny będzie crawler, zaimplementowany w języku C# przy użyciu biblioteki HtmlAgilityPack służącej do pobierania i parsowania stron internetowych. Dane będą zapisywane w relacyjnej bazie danych Microsoft SQL, której schemat opisany jest w kolejnych punktach. Crawler zaczynać będzie swoje działanie od strony na angielskiej wikipedii prezentującej listę wszystkich bitew toczonych w latach 1301-1800.¹ Następnie, będzie odwiedzał strony kolejnych bitew, wyciągając z nich i zapisując w bazie danych informacje (o bitwach, stronach konfliktu i liderach). Dla każdej bitwy crawler zapisze sobie informację na temat wojny do której ona należy, aby po zakończeniu przeglądania bitew dysponować listą wojen wraz z adresami URL, co umożliwi crawling wojen w analogiczny sposób jak bitew. Po zakończeniu pobierania bitew i wojen, crawler przejdzie do strony z listą pokojów i również pobierze o nich informacje. Jako że tabele podsumowujące informacje o bitwach i wojnach są na angielskiej wikipedii w przeważającej części ustandaryzowane, crawling tych informacji jest stosunkowo łatwe. Większy problem stanowią pokoje, dla których nie jest ustalony żaden standard, w związku z czym w łatwy sposób można pobrać jedynie datę, nazwę i podsumowanie pokoju.

PRZETWARZANIE DANYCH

Po pozyskaniu danych przez crawler, występują one w przeważającej części w postaci pól tekstowych, co praktycznie uniemożliwia automatyczną analizę za pomocą

¹ https://en.wikipedia.org/wiki/List_of_battles_1301-1800

dostępnych nam narzędzi. W związku z tym drugim etapem musi być przetworzenie danych tak, aby pola tekstowe przetransformować na odpowiednie pola liczbowe (np. w przypadku rozmiaru sił uczestniczących w bitwie), lub też na jednoznaczne (enumerowane) pola tekstowe (np. w przypadku wyniku wojny). Aby uzyskać takie dane, niezbędne będzie zaimplementowanie odpowiednich parserów, wyciągających kluczowe informacje z odpowiednich pól tekstowych. Przy implementacji zastosowane będą przede wszystkim wyrażenia regularne i słowniki.

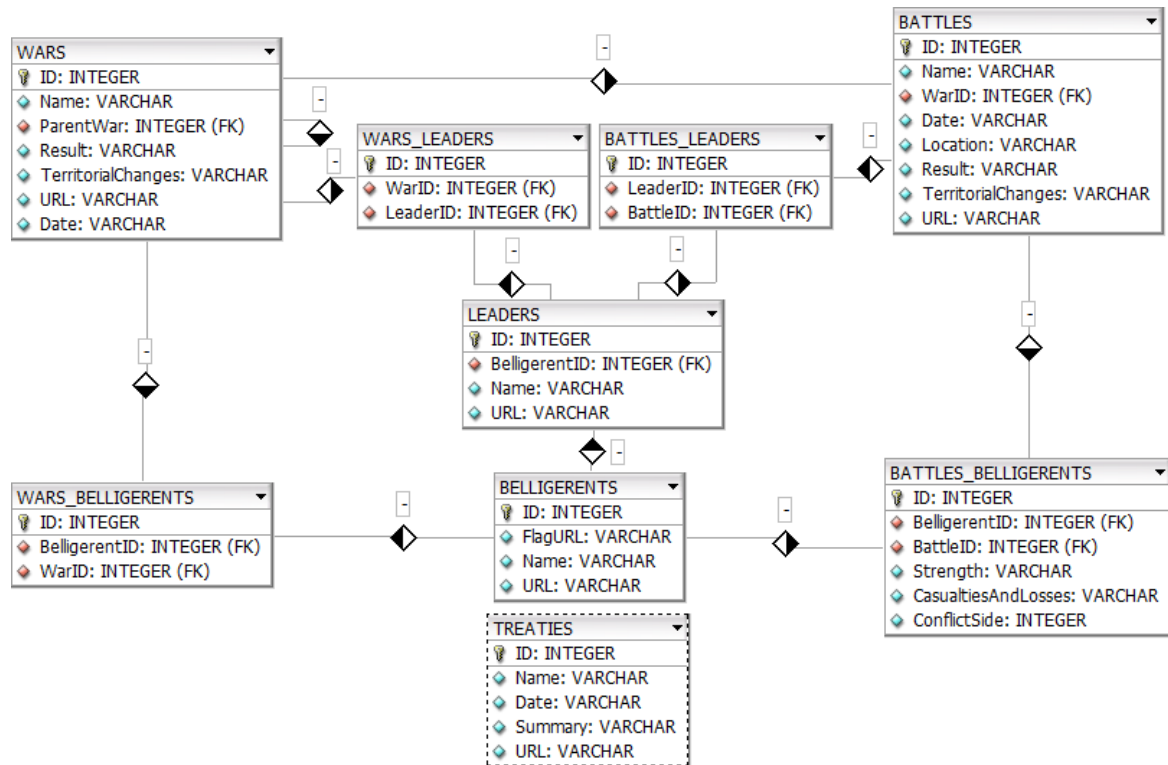
ANALIZA I WIZUALIZACJA DANYCH

Mając już pobrane i przetworzone w odpowiedni sposób dane, będzie można przejść do części ostatniej, czyli analizy danych. Będzie ona podzielona na dwie części. W pierwszej skupimy się na wyciągnięciu ciekawych statystyk za pomocą bezpośrednich zapytań SQL (w ten sposób można np. odkryć pewne trendy, najbardziej zaangażowane w walki państwa, czy też zmienianie się pewnych statystyk w czasie). Uzyskane informacje zamierzamy w większości zwizualizować w postaci wykresów. Drugą częścią będzie analiza sieci (państw, dowódców) korzystając z odpowiednich do tego narzędzi (Gephi). Wizualizacja będzie polegała na przedstawieniu grafów z wyróżnieniem najistotniejszych węzłów.

4. Baza danych

W projekcie zastosowana została relacyjna baza danych Microsoft SQL.

Podstawowy schemat bazy (zawierający wyłącznie surowe dane z crawlowania, bez kolumn przetworzonych):



5. Szczegółowy opis tabel

TABELA WARS

Reprezentuje wojny.

- **ID** - *PK, int, not null* - identyfikator wojny, klucz główny
- **Name** - *nvarchar, null* - nazwa wojny
- **ParentWar** - *FK, int, null* - identyfikator głównej wojny, której częścią jest dana wojna, klucz obcy do tabeli *WARS*
- **Result** - *nvarchar, null* - wynik, jakim zakończyła się wojna
- **TerritorialChanges** - *nvarchar, null* - zmiany terytorialne do których doszło w wyniku wojny
- **URL** - *nvarchar, null* - adres URL pod którym znajduje się strona internetowa wojny na angielskiej wikipedii
- **Date** - *nvarchar, null* - data odbycia się wojny

TABELA BATTLES

Reprezentuje bitwy.

- **ID** - *PK, int, not null* - identyfikator bitwy, klucz główny
- **Name** - *nvarchar, null* - nazwa bitwy
- **WarID** - *FK, int, null* - identyfikator wojny w ramach której odbywała się bitwa, klucz obcy do tabeli *WARS*
- **Date** - *nvarchar, null* - data odbycia się bitwy
- **Location** - *varchar, null* - miejsce w którym odbyła się bitwa
- **Result** - *varchar, null* - wynik jakim zakończyła się bitwa
- **TerritorialChanges** - *varchar, null* - zmiany terytorialne do których doszło w wyniku bitwy
- **URL** - *varchar, null* - adres URL pod którym znajduje się strona internetowa bitwy na angielskiej wikipedii

TABELA BELLIGERENTS

Reprezentuje strony konfliktów.

- **ID** - *PK, int, not null* - identyfikator strony konfliktu, klucz główny

- **FlagURL** - *varchar, null* - adres URL pod którym znajduje się obrazek z flagą reprezentującą stronę konfliktu
- **Name** - *varchar, null* - nazwa strony konfliktu
- **URL** - *varchar, null* - adres URL pod którym znajduje się strona internetowa strony konfliktu na angielskiej wikipedii

TABELA WARS_BELLIGERENTS

Tabela łącznikowa łącząca tabele *WARS* i *BELLIGERENTS* w relacji wiele do wielu. Reprezentuje przypisanie stron konfliktu do wojen.

- **ID** - *PK, int, not null* - identyfikator relacji, klucz główny
- **BelligerentID** - *FK, int, null* - identyfikator strony konfliktu, klucz obcy do tabeli *BELLIGERENTS*
- **WarID** - *FK, int, null* - identyfikator wojny, klucz obcy do tabeli *WARS*

TABELA BATTLES_BELLIGERENTS

Tabela łącznikowa łącząca tabele *BATTLES* i *BELLIGERENTS* w relacji wiele do wielu. Reprezentuje przypisanie stron konfliktu do bitew.

- **ID** - *PK, int, not null* - identyfikator relacji, klucz główny
- **BelligerentID** - *FK, int, null* - identyfikator strony konfliktu, klucz obcy do tabeli *BELLIGERENTS*
- **BattleID** - *FK, int, null* - identyfikator bitwy, klucz obcy do tabeli *BATTLES*
- **Strength** - *varchar, null* - siły strony konfliktu uczestniczące w bitwie
- **CasualtiesAndLosses** - *varchar, null* - straty strony konfliktu w czasie bitwy
- **ConflictSide** - *varchar, null* - strona po której opowiadali się walczący w czasie bitwy ("lewa" i "prawa" strona na wikipedii)

TABELA LEADERS

Reprezentuje dowódców wojennych.

- **ID** - *PK, int, not null* - identyfikator dowódcy, klucz główny
- **BelligerentID** - *FK, int, null* - identyfikator strony konfliktu dowódcy, klucz obcy do tabeli *BELLIGERENTS*
- **Name** - *varchar, null* - imię i nazwisko dowódcy

- **URL** - *varchar, null* - adres URL pod którym znajduje się strona internetowa dowódcy na angielskiej wikipedii

TABELA WARS_LEADERS

Tabela łącznikowa łącząca tabele *WARS* i *LEADERS* w relacji wiele do wielu. Reprezentuje przypisanie dowódców do wojen.

- **ID** - *PK, int, not null* - identyfikator relacji, klucz główny
- **WarID** - *FK, int, null* - identyfikator wojny, klucz obcy do tabeli *WARS*
- **LeaderID** - *FK, int, null* - identyfikator dowódcy, klucz obcy do tabeli *LEADERS*

TABELA BATTLES_LEADERS

Tabela łącznikowa łącząca tabele *BATTLES* i *LEADERS* w relacji wiele do wielu. Reprezentuje przypisanie dowódców do bitew.

- **ID** - *PK, int, not null* - identyfikator relacji, klucz główny
- **LeaderID** - *FK, int, null* - identyfikator dowódcy, klucz obcy do tabeli *LEADERS*
- **BattleID** - *FK, int, null* - identyfikator bitwy, klucz obcy do tabeli *BATTLES*

TABELA TREATIES





Reprezentuje pokoje.

- **ID** - *PK, int, not null* - identyfikator pokoju, klucz główny
- **Name** - *varchar, null* - nazwa pokoju
- **Date** - *date, null* - data zawarcia pokoju
- **Summary** - *varchar, null* - tekstowe podsumowanie warunków pokoju
- **URL** - *varchar, null* - adres URL pod którym znajduje się strona internetowa pokoju na angielskiej wikipedii






6. Mapowanie danych źródłowych

Mapowanie danych źródłowych pochodzących ze stron bitew i wojen na angielskiej wikipedii na kolumny w bazie danych.

Bitwy:

Battle of Klushino	[BATTLES].NAME
Part of Polish–Russian War (1605–1618)	[WARS].NAME
	
Polish <i>hussar</i> line at the Battle of Klushyn	
Date 4 July 1610	[BATTLES].DATE
Location Klushino	[BATTLES].LOCATION
Result Decisive Polish victory	[BATTLES].RESULT
Belligerents	[BELLIGERENTS].FLAGURL
 Polish-Lithuanian Commonwealth	
 Tsardom of Russia	
 Sweden	[BELLIGERENTS].NAME
Commanders and leaders	
Stanisław Żółkiewski	
Dmitry Shuisky	[LEADERS].NAME
Jacob De la Gardie	
Strength	
6,500–6,800 men ^{[1][2]}	
30,000 Russians ^{[2][3]} and 5,000 mercenaries	[BATTLES_BELLIGERENTS].STRENGTH
2 guns ^[2]	
11 guns ^[2]	
Casualties and losses	
400 ^[2]	
5,000 ^[2]	[BATTLES_BELLIGERENTS].CASUALTIESANDLOSSES

Wojny:

Polish–Russian War of 1605–1618		[WARS].NAME
 <p>Map of the war. Important battles marked with crossed swords.</p>		
Date	1605–1618	[WARS].DATE
Location	Russia	[WARS].LOCATION
Result	<ul style="list-style-type: none"> Polish victory in gaining territory; however, failure in destroying the Russian state. Russia retained independence. Truce of Deulino 	[WARS].RESULT
Belligerents		
 Polish–Lithuanian Commonwealth	 Tsardom of Russia	[BELLIGERENTS].FLAGURL
	 Kingdom of Sweden (1609–1610)	[BELLIGERENTS].NAME
Commanders and leaders		
King Sigismund III King Władysław IV	 Boris Godunov Mikhail Skopin-Shuysky Jakob De la Gardie Dmitry Pozharsky	[LEADERS].NAME

7. Przetwarzanie pobranych danych

Po etapie pobrania danych, tabele zawierają komplet informacji o bitwach i wojnach, natomiast dane w nich nie są odpowiednio przetworzone. Zamiast pól tekstowych, po etapie przetwarzania danych, powinniśmy otrzymać dane liczbowe uszeregowane względem konkretnych wymiarów. Operacje, które powinny zostać wykonane podczas przetwarzania, zostały wypisane poniżej.

a. Parsowanie pól tekstowych

- Tabela BATTLES:
 - Date - parsowanie w celu uzyskania daty w odpowiednim formacie
 - Location - parsowanie w celu uzyskania państwa, na terenie którego odbyła się bitwa
 - Result - parsowanie w celu uzyskania zwycięzcy
- Tabela WARS:
 - Date - parsowanie w celu uzyskania daty w odpowiednim formacie
 - Result - parsowanie w celu uzyskania zwycięzcy, w trudniejszym wariancie w celu informacji o utraconych/zdobytych ziemiach
- Tabela BATTLES_BELLIGERENTS:
 - Strength - parsowanie w celu uzyskania dokładnego rozkładu sił pośród sojuszników
 - CasualtiesAndLosses - parsowanie w celu uzyskania dokładnego rozkładu strat pośród sojuszników
- Tabela BELLIGERENTS:
 - Name - parsowanie w celu pogrupowania tych samych frakcji w jedno, np. **Kingdom of France** i **France**
- Tabela TREATIES:
 - Summary - parsowanie w celu uzyskania relacji między pokojami a wojnami/bitwami
- b. Uzupełnianie danych - jeżeli crawler zawiedzie dla części danych, zostaną one ręcznie przetworzone (dodane lub usunięte w zależności od konkretnych przypadków).

Pola tekstowe TerritorialChanges w BATTLES i WARS nie zostaną przetworzone, gdyż w naszej analizie nie zajmujemy się tym aspektem.

8. Modyfikacja modelu

W związku z koniecznością zapisania stanu przetworzonych danych, wprowadzone zostały modyfikacje do schematu modelu bazy danych:

- Tabela BATTLES:
 - **StartDate** - *date, null* - data rozpoczęcia się bitwy
 - **EndDate** - *date, null* - data zakończenia bitwy
 - **Country** - *nvarchar, null* - dzisiejsze państwo na terenie którego odbyła się bitwa
- Tabela WARS:
 - **StartDate** - *date, null* - data rozpoczęcia się wojny
 - **EndDate** - *date, null* - data zakończenia wojny
- Tabela BATTLES_BELLIGERENTS:
 - **Result** - *nvarchar, null* - rezultat bitwy dla danej strony; enumerator
 - **InfantryStrength** - *int, null* - sumaryczna liczba piechoty
 - **CavalryStrength** - *int, null* - sumaryczna liczba kawalerii
 - **ArtilleryStrength** - *int, null* - sumaryczna liczba artylerii
 - **NavyStrength** - *int, null* - sumaryczna liczba statków
 - **AllStrength** - *int, null* - sumaryczna liczba piechoty i artylerii
 - **OtherStrength** - *nvarchar, null* - ilość żołnierzy jeśli nie podana liczbowo; enumerator
 - **Killed** - *int, null* - liczba osób zabitych w bitwie
 - **Wounded** - *int, null* - liczba osób rannych w bitwie
 - **Captured** - *int, null* - liczba osób w niewoli po bitwie
 - **AllLosses** - *int, null* - sumaryczna liczba strat w bitwie
 - **OtherLosses** - *nvarchar, null* - straty w bitwie dla danej strony jeśli nie podane liczbowo; enumerator
- Tabela WARS_BELLIGERENTS:
 - **Result** - *nvarchar, null* - rezultat wojny dla danej strony; enumerator.
- Tabela BELLIGERENTS:
 - **CountryGroup** - *nvarchar, null* - grupa do której należy dana frakcja

9. Sposób przetworzenia danych

Podstawowym narzędziem do wydobywania danych potrzebnych do dalszych analiz jest dość ogólny parser, którego implementacja jest dostosowana do tekstowego formatu danych źródłowych, a także do formatu docelowego. Poniżej opisana została strategia działania parsera oraz podstawowe założenia dla przetwarzania kolejnych pól tekstowych.

Ogólne założenia:

- Wszelkie znaki specjalne a także zawartość nawiasów dowolnych rodzajów jest usuwana przed analizą.

a. Parsowanie liczebności sił (**Strength**):

Parser wykrywa słowa kluczowe w tekście:

Infantry, Artillery, Cavalry, Ships, Soldier, Archer, Gun, Cannon, Regular, Militia

Liczba bezpośrednio przed słowem kluczowym traktowana jest jako liczebność danego typu sił. Jeśli jest to przedział, brana jest pod uwagę średnia. Jeśli w tekście mamy tylko liczbę, to dane trafiają do kolumny **AllStrength**.

Skuteczność działania parsera: około **60%**

Problemy przy przetwarzaniu: estymaty sił zamiast konkretnych danych, nazwy oddziałów lub specyficznych jednostek niemożliwe do rozdysponowania, różne źródła liczebności sił, różne fazy bitwy (w każdej fazie różne liczby żołnierzy)

Weryfikacja danych: pozostałe **40%** zostało przetworzonych manualnie.

b. Parsowanie liczebności strat (**CasualtiesAndLosses**):

Parser wykrywa słowa kluczowe w tekście:

Killed, Wounded, Captured, Imprisoned, Missing, Casualties, Lost, Sunk

Liczba bezpośrednio przed słowem kluczowym traktowana jest jako liczebność danego typu strat. Jeśli jest to przedział, brana jest pod uwagę średnia. Jeśli w tekście mamy tylko liczbę, to dane trafiają do kolumny **Killed**. Jeśli liczba dotyczy określenia Casualties, lub podana jest jedna liczba na określenie zabitych i schwytanych, to dane trafiają do kolumny **AllLosses**. Jeśli jednak jedna liczba dotyczy zabitych i rannych, to te dane trafiają do kolumny **Killed**. Dodatkowo jeśli opis strat jest dany przymiotnikiem: **Heavy, Very Heavy, Hard, Very Hard, Light, Very Light, Few, None, Total** to taką informację zapisujemy w kolumnie

OtherLosses jako jedną z 6 wartości predefiniowanych: None, Very Light, Light, Heavy, Very Heavy, Total

Skuteczność działania parsera: **70%**

Problemy przy przetwarzaniu: sumowanie częściowe strat bez możliwości rozdzielenia na mniejsze (np. 2 tys. rannych i zaginionych), estymaty strat zamiast konkretnych danych, różne źródła liczebności strat, straty podane osobno dla całej bitwy oraz następstw tej bitwy, straty uwzględniające cywilów lub nie.

Weryfikacja danych: pozostałe **30%** zostało przetworzonych manualnie.

c. Parsowanie dat (**StartDate,EndDate**)

Parser próbuje konwertować literał tekstowy do klasy DateTime. W przypadku istnienia różnych dat początkowej i końcowej, parser używa funkcji split, oddzielając według znaku '-' obie daty i dokonując ich osobnej analizy.

Skuteczność działania parsera: **70%**

Problemy przy przetwarzaniu: nieregularne formaty dat, podawanie kilku dat zamiast jednej, słowne opisy zawierające np pory roku zamiast miesięcy. Ze względu na ograniczenie typu date w SQL Server (datowanie od 1732 roku), przed wprowadzeniem daty do bazy do każdej daty dodawane jest 1000 lat. Jeśli podawany jest tylko rok, albo miesiąc i rok, data automatycznie ustalana jest jako 1 stycznia danego roku.

Weryfikacja danych: Data startowa musi być wcześniejsza niż końcowa, a także jeśli jedna z nich jest ustawiona na 1 stycznia, druga także jest nadpisywana.

Daty spoza przedziału są usuwane.

d. Parsowanie krajów (**Location -> Country**)

Parser dysponuje predefiniowaną listą krajów, i szuka w kolumnie **Location** ich wystąpień.

Skuteczność działania parsera: **55%**

Problemy przy przetwarzaniu: nazwy regionów/stanów (np:California) lub miast zamiast nazw państw, używanie nazw historycznych, jedynie współrzędne

geograficzne, bitwy morskie - brak przynależności do konkretnego państwa

Weryfikacja danych: brak.

e. Parsowanie rezultatów bitew (**Result**)

Parser wykrywa słowa kluczowe w tekście: **Decisive Victory/Victory** a następnie zapisuje literał występujący po tych słowach. Dysponując nazwami wszystkich frakcji w danej bitwie, parser dopasowuje do nich litera po literze znaleziony literał - decyduje dłuższe dopasowanie. Ostatecznie w docelowej kolumnie znajduje się jedna z 6 wartości: **Decisive Win, Win, Lose, Decisive Lose, Inconclusive** (w przypadku braku słowa Victory) lub **NULL** (niepowodzenie parsera).

Skuteczność działania parsera: **60%**

Problemy przy przetwarzaniu: formalizmy - z tej przyczyny przy analizie dopasowań z nazw frakcji usuwane są słowa kluczowe, m.in.: **Kingdom of, County of, Republic of, Empire of** itp.; specyficzne odmiany przymiotników w języku angielskim (**Netherlands -> Dutch**), używanie synonimów (**Ottomans - Turkey, Catalonia - Majorca**)

Weryfikacja danych: pozostałe dane przetworzone manualnie, zgodnie z podstawową wiedzą.

f. Parsowanie grup frakcji (**Name -> GroupCountry**)

Brak możliwości parsowania ze względu na brak informacji w danych źródłowych.

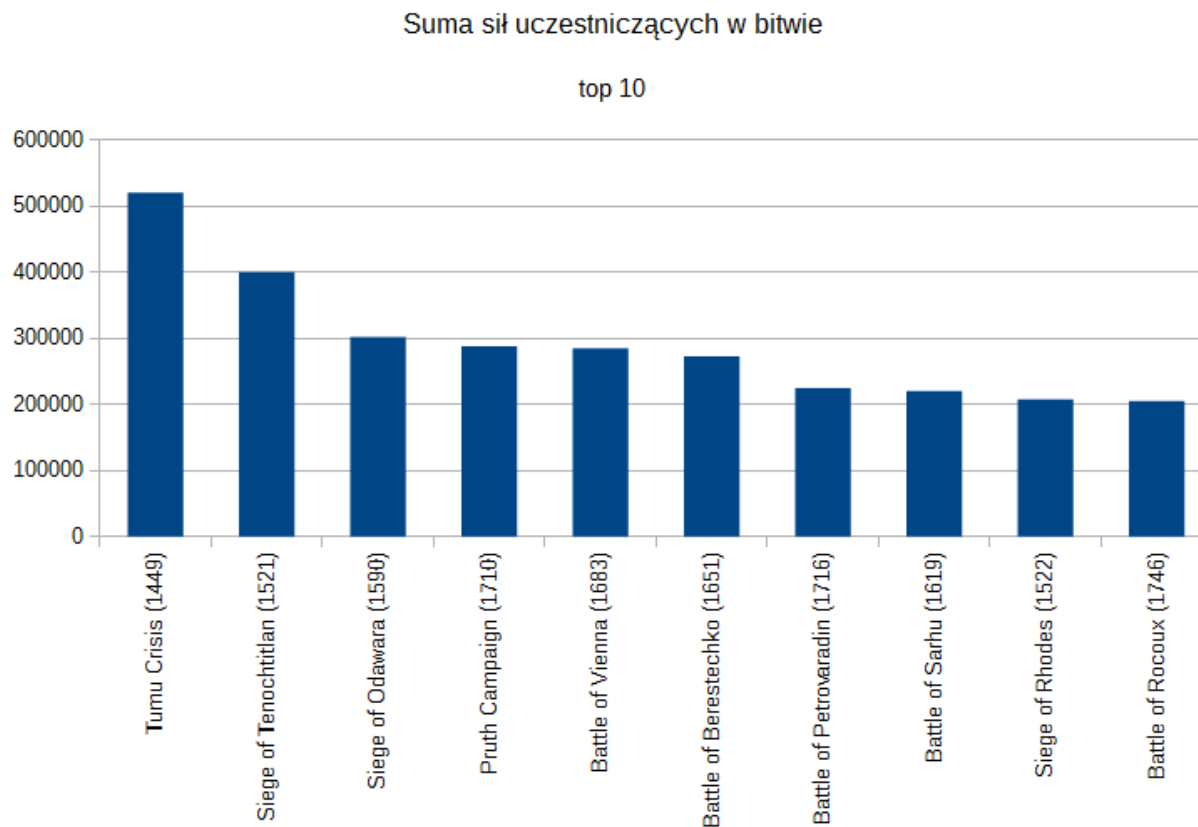
Rozwiązanie: przyporządkowanie manualne części frakcji do jednej grupy na podstawie elementarnej wiedzy historycznej.

10. Wizualizacja i analiza danych

a. Podstawowe wykresy

Spis wykresów:

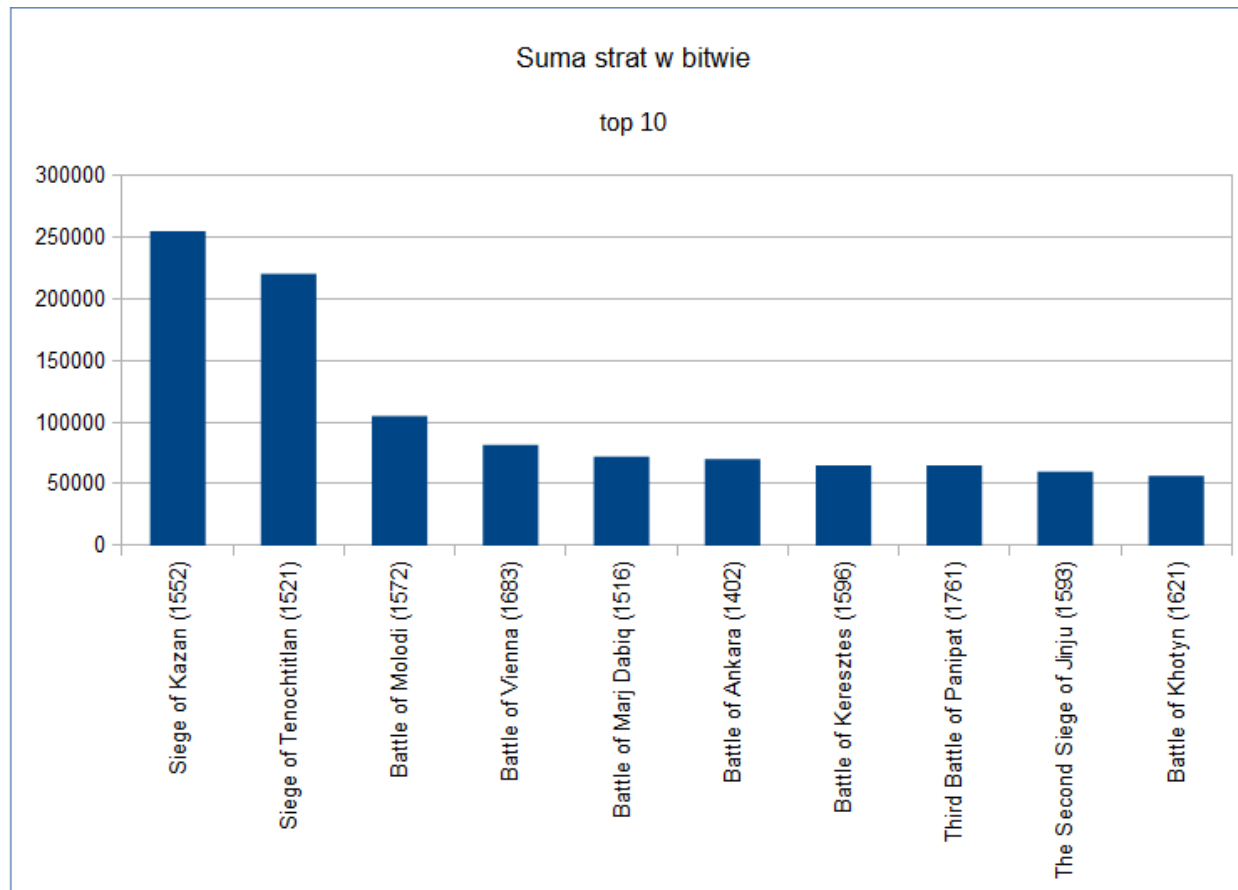
- Wykres nr 1: Największe bitwy wg sumy sił w danej bitwie.
- Wykres nr 2: Największe bitwy wg sumy strat w danej bitwie.
- Wykres nr 3: Bitwy z największym stosunkiem strat do sił.
- Wykres nr 4: Dzisiejsze państwa, na terenie których odbyło się najwięcej bitew.
- Wykres nr 5: Największe wojny wg sumy sił w bitwach.
- Wykres nr 6: Największe wojny wg liczby bitew.
- Wykres nr 7: Strony konfliktu z największą liczbą zwycięstw.
- Wykres nr 8: Strony konfliktu z największą procentową ilością zwycięstw w bitwach (uczestniczący w co najmniej 10 bitwach).
- Wykres nr 9: Strony konfliktu z największą procentową ilością zwycięstw w bitwach bez uwzględniania tych nierozstrzygniętych(uczestniczący w co najmniej 10 bitwach).
- Wykres nr 10: Strony konfliktu z największą ilością stoczonych bitew.
- Wykres nr 11: Liderzy z największą liczbą stoczonych bitew.
- Wykres nr 12: Liczba bitew na przestrzeni lat (grupowane co 20 lat).
- Wykres nr 13: Liczba sił na przestrzeni lat (grupowane co 20 lat).
- Wykres nr 14: Liczba strat na przestrzeni lat (grupowane co 20 lat).
- Wykres nr 15: Podział sił na typy jednostek na przestrzeni lat (grupowane co 20 lat).
- Wykres nr 16: Ile razy zwyciężały strony konfliktu mające przeważające siły w bitwach.
- Wykres nr 17: Ile razy strony konfliktu z przeważającymi siłami ponosiły większe straty.

Wykres nr 1: Największe bitwy wg sumy sił w danej bitwie.

Bitwa	Suma sił
Tumu Crisis (1449)	520000
Siege of Tenochtitlan (1521)	400100
Siege of Odawara (1590)	302000
Pruth Campaign (1710)	288000
Battle of Vienna (1683)	285000
Battle of Berestechko (1651)	273000
Battle of Petrovaradin (1716)	225000
Battle of Sarhu (1619)	220000
Siege of Rhodes (1522)	207500
Battle of Rocoux (1746)	205000
Bătălia de la Vaslui (1475)	204020
Battle of Szigeth (1566)	202500
Battle of Sprimont (1794)	199000
Battle of Warburg (1760)	192000
Siege of Kazan (1552)	190000
Battle of Molodi (1572)	185000
Battle of Sekigahara (1600)	180000
Battle of Kulikovo (1380)	180000
Battle of Khotyn (1621)	178000
Battle of Cahul (1770)	178000

Wykres wskazuje na duży udział oblężień. Dzieje się tak dlatego, że w źródłach często doliczano mieszkańców oblężanego miasta do sumy sił walczących w bitwie. Oblężenie azteckiej stolicy było nieco innym przypadkiem, gdyż Indian-wojowników naliczono aż 300 000. Warto też zwrócić uwagę, że bitwy nie są skupione w jednym okresie, ani nie mają wspólnego podłoża geopolitycznego.

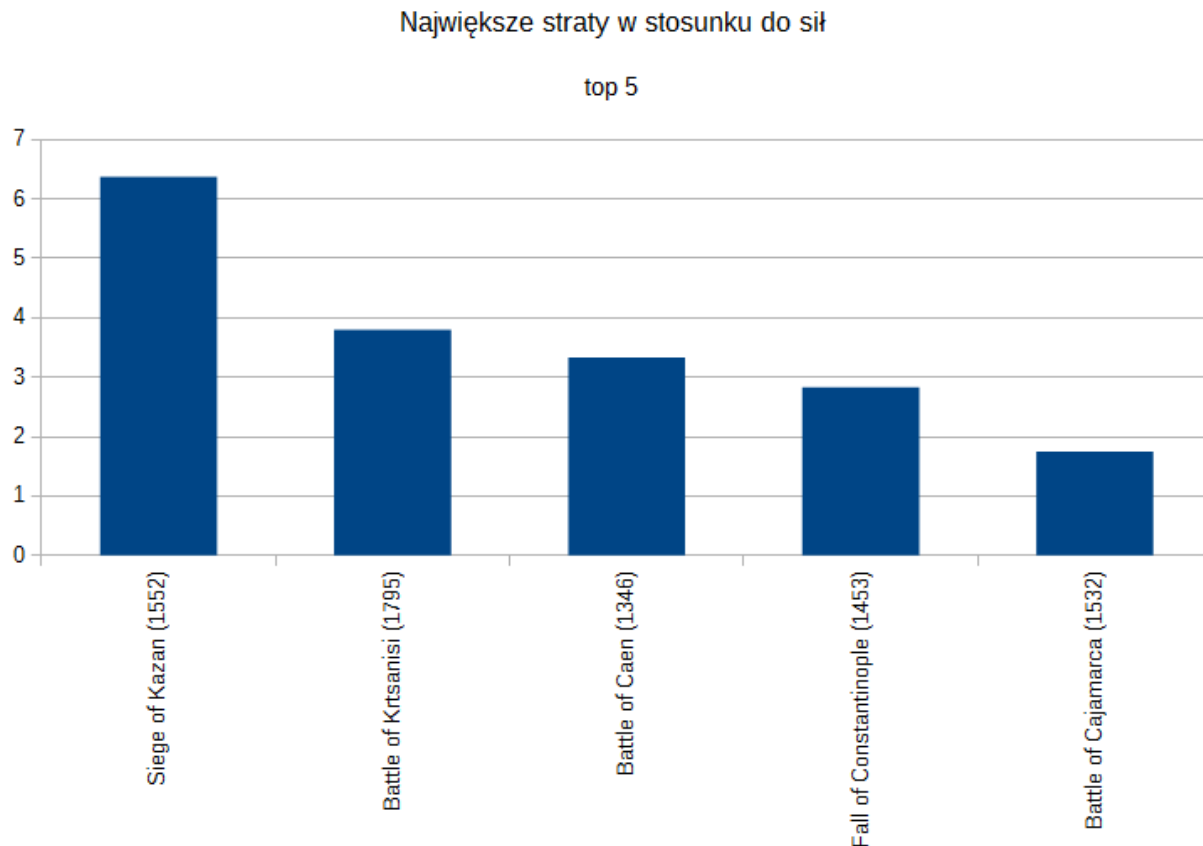
Wykres nr 2: Największe bitwy wg sumy strat w danej bitwie.



Bitwa	Suma sił
Siege of Kazan (1552)	255000
Siege of Tenochtitlan (1521)	220400
Battle of Molodi (1572)	105000
Battle of Vienna (1683)	81500
Battle of Marj Dabiq (1516)	72000
Battle of Ankara (1402)	70000
Battle of Keresztes (1596)	65000
Third Battle of Panipat (1761)	65000
The Second Siege of Jinju (1593)	60000
Battle of Khotyn (1621)	56500
Battle of Chungju (Choryang Pass) (1592)	50000
Battle of Kunersdorf (1759)	49574
Battle of Sarhu (1619)	47000
Battle of Blenheim (1704)	46674
Battle of Berestechko (1651)	35700
Battle of Varna (1444)	35000
Battle of Petrovaradin (1716)	35000
Battle of Chudnov (Cudnów) (1660)	34900
Battle of Grunwald (1410)	34800
Fall of Constantinople (1453)	34000

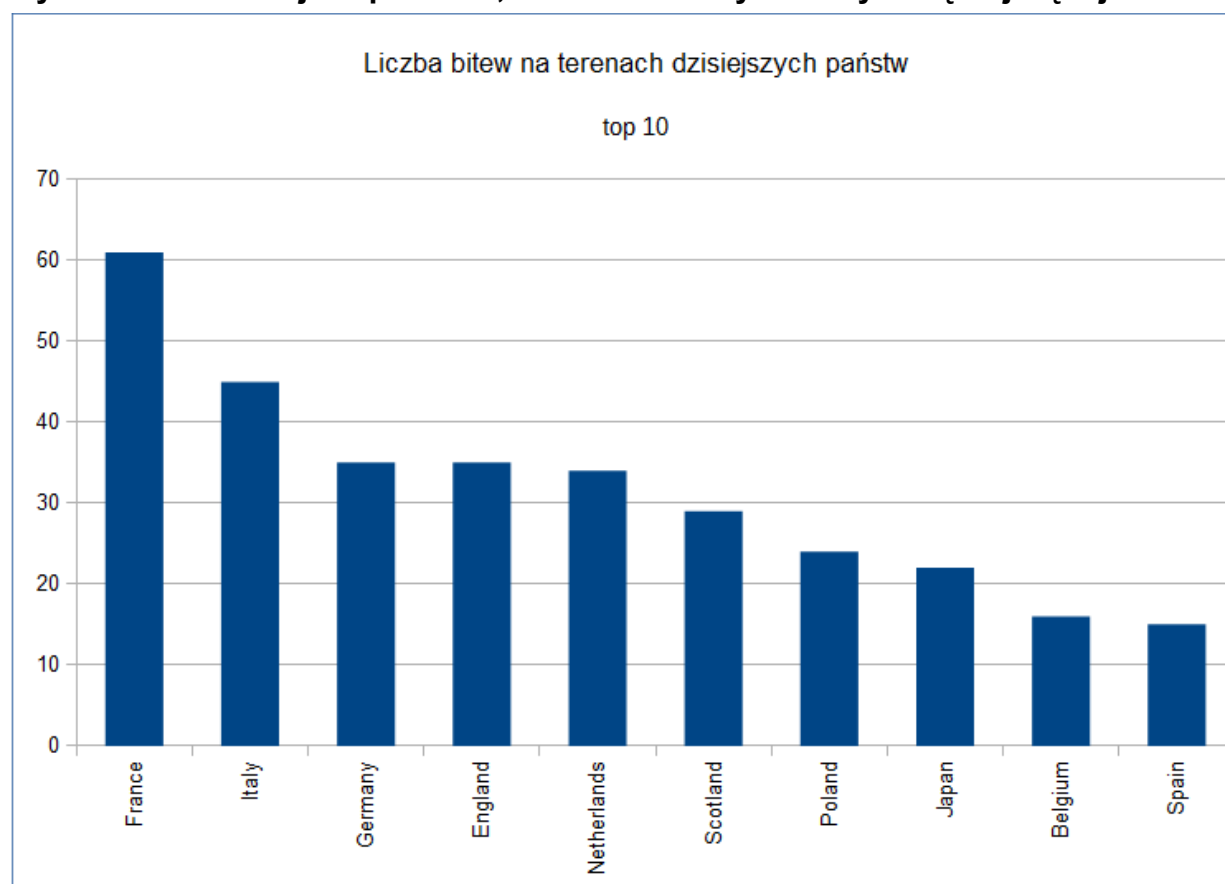
W przypadku wykresu nr 2 widać, że zdarzało się że straty w bitwie przekraczały całkowitą sumę żołnierzy w niej walczących (Siege of Kazan, 1552). Przyczyną jest niejednorodność podawanych danych i udział cywilów w stratach (którzy nie byli uwzględniani w siłach). Widać również, że bitwy przodujące na wykresie miały miejsce na wschodzie, gdzie technologia była słabsza i stawiano na ilość a nie na jakość.

Wykres nr 3: Bitwy z największym stosunkiem strat do sił.



Wykres nr 3 wskazuje niemal wyłącznie na niekonsekwencję w danych źródłowych - straty w bitwach są wielokrotnie wyższe niż siły które brały w nich udział. Przyczyną takiego stanu rzeczy, oprócz udziału cywilów w stratach, były dane źródłowe wskazujące na często wykluczające się siły i straty obu stron, z których każda umniejszała liczebność swoich sił a zwiększała liczebność wroga.

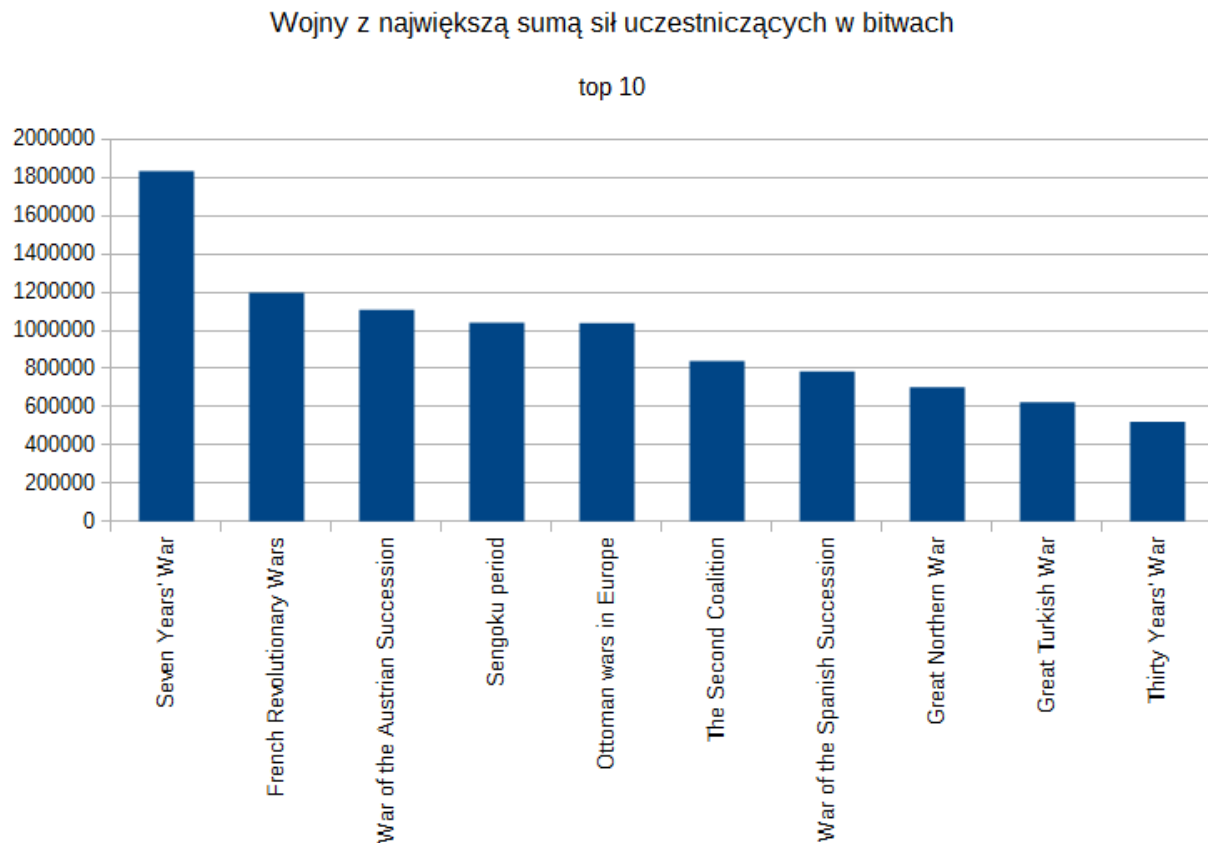
Wykres nr 4: Dzisiejsze państwa, na terenie których odbyło się najwięcej bitew.



Państwo	Liczba bitew
France	61
Italy	45
Germany	35
England	35
Netherlands	34
Scotland	29
Poland	24
Japan	22
Belgium	16
Spain	15
India	14
Czech Republic	11
Portugal	11
Ukraine	10
Serbia	8
Ethiopia	8
Hungary	7
Korea	7
Brazil	6
Peru	6

Wykres wskazuje na miejsca, gdzie w latach 1300-1800 toczyło się najwięcej konfliktów zbrojnych. Rezultaty nie są niespodziankami, natomiast ciekawy jest duży udział Szkocji i państw Beneluksu, a mały Austrii, Hiszpanii czy państw bałkańskich. Nieobecność Stanów Zjednoczonych związana jest z niedokładnością parsera, w źródłach pojawiał się jedynie stan USA a nie państwo.

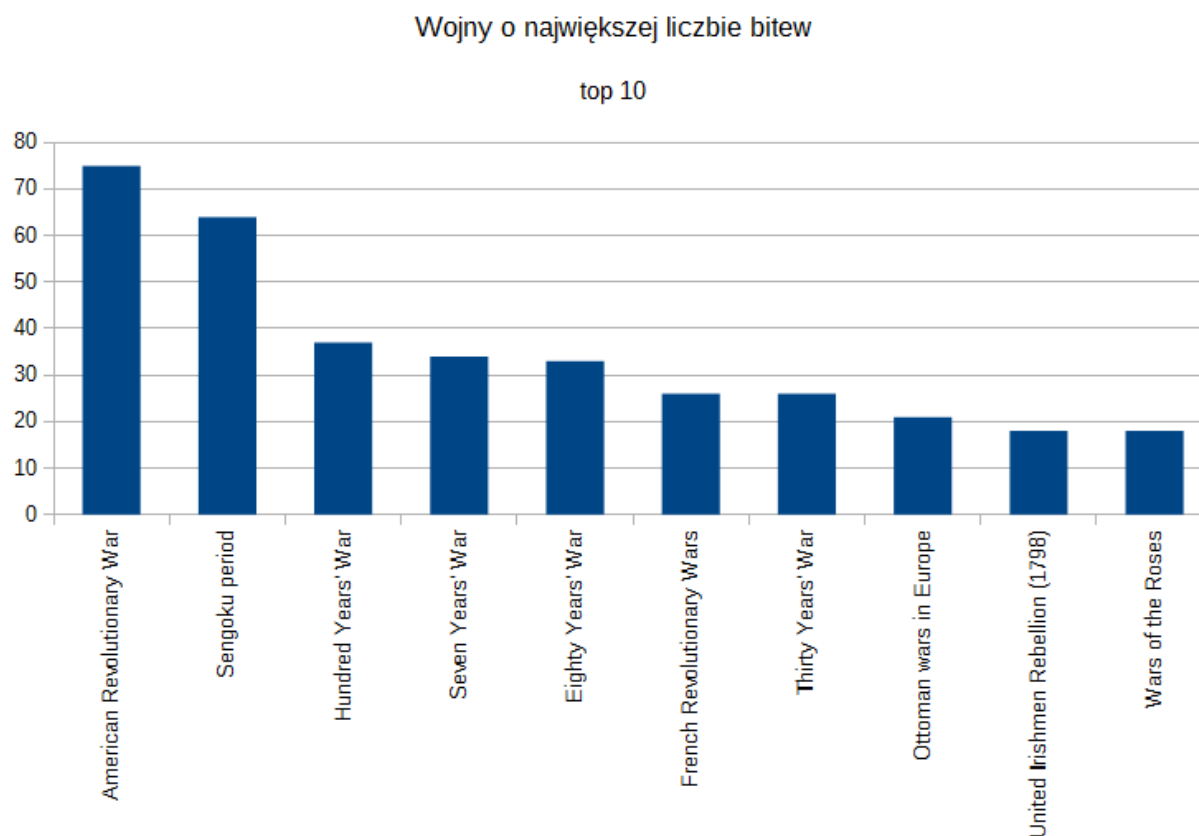
Wykres nr 5: Największe wojny wg sumy sił w bitwach.



Wojna	Suma sił
Seven Years' War	1833500
French Revolutionary Wars	1198008
War of the Austrian Succession	1107280
Sengoku period	1040200
Ottoman wars in Europe	1039000
The Second Coalition	839598
War of the Spanish Succession	786200
Great Northern War	701841
Great Turkish War	625000
Thirty Years' War	521558
Hussite Wars	495350
Hundred Years' War	488450
Nine Years' War	477306
Japanese invasions of Korea	474600
Spanish conquest of the Aztec Empire	471100
Franco–Dutch War	455057
Eighty Years' War	391345
Moldavian–Ottoman Wars	354020
War of the League of Cambrai	344100
Khmelnitsky Uprising	273000

Największymi wojnami okazały się wojna siedmioletnia, wojna związana z francuską rewolucją (1792) i wojna o sukcesję austriacką. Co ciekawe, wszystkie te wojny były w XVIII wieku a żadna z tych wojen nie miała olbrzymich bitew jak we wcześniejszych latach. Były one również znacznie krótsze niż, przykładowo, wojna stuletnia czy trzydziestoletnia. Z drugiej strony były znacznie bardziej globalnymi konfliktami na skalę światową i często walki toczyły się na wielu kontynentach.

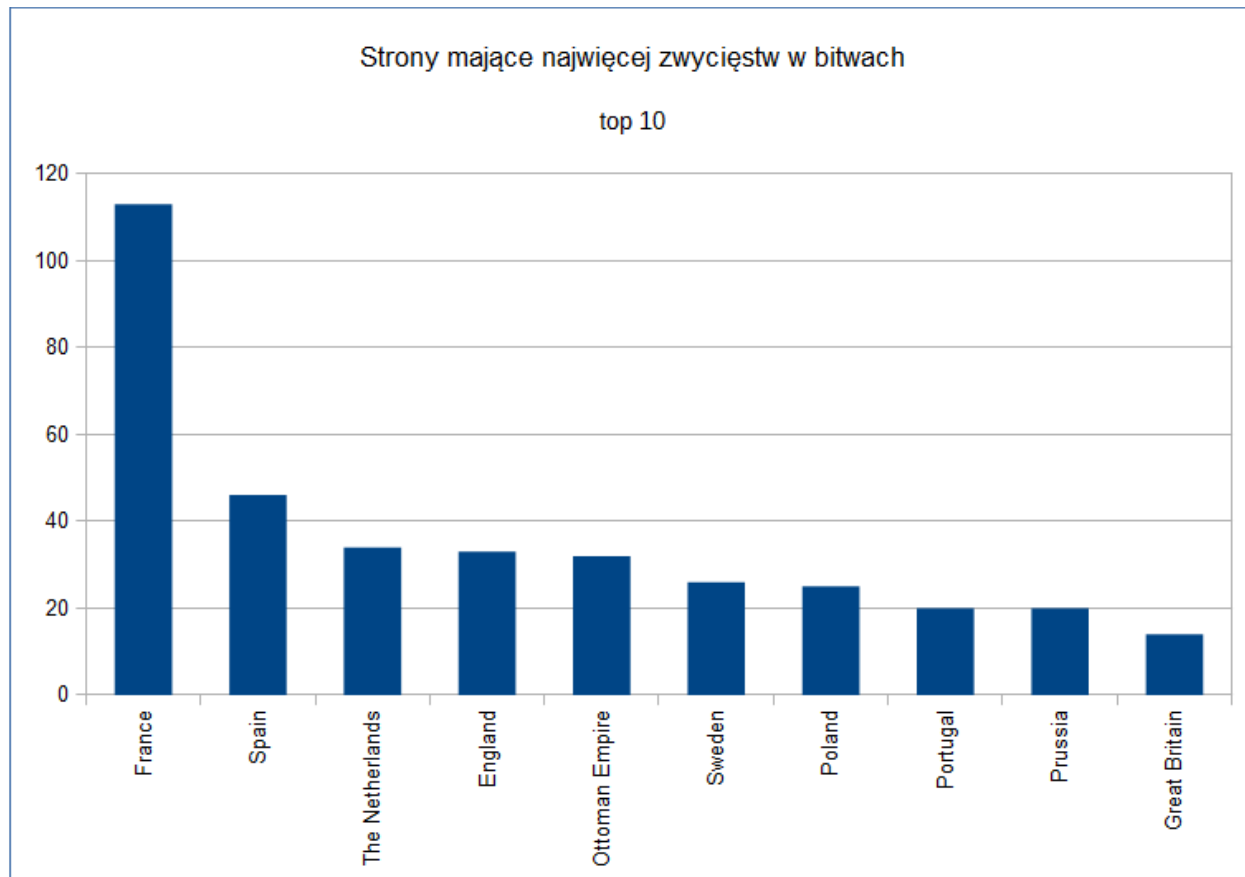
Wykres nr 6: Największe wojny wg liczby bitew.



Wojna	Liczba bitew
American Revolutionary War	75
Sengoku period	64
Hundred Years' War	37
Seven Years' War	34
Eighty Years' War	33
French Revolutionary Wars	26
Thirty Years' War	26
Ottoman wars in Europe	21
United Irishmen Rebellion (1798)	18
Wars of the Roses	18
War of the Spanish Succession	17
War of the Austrian Succession	17
Great Northern War	17
Japanese invasions of Korea	16
French and Indian War	16
Dutch-Portuguese War	15
Franco-Dutch War	14
Nine Years' War	13
English Civil War	12
Abyssinian- Adal war	9

Duża liczba bitew w rewolucji amerykańskiej wynika z wielkiej ilości małych potyczek, które przez "historyków" z angielskiej wikipedii zostały podniesione do rangi bitew. Widać też dominację dłuższych wojen nad krótszymi w dalszej części zestawienia.

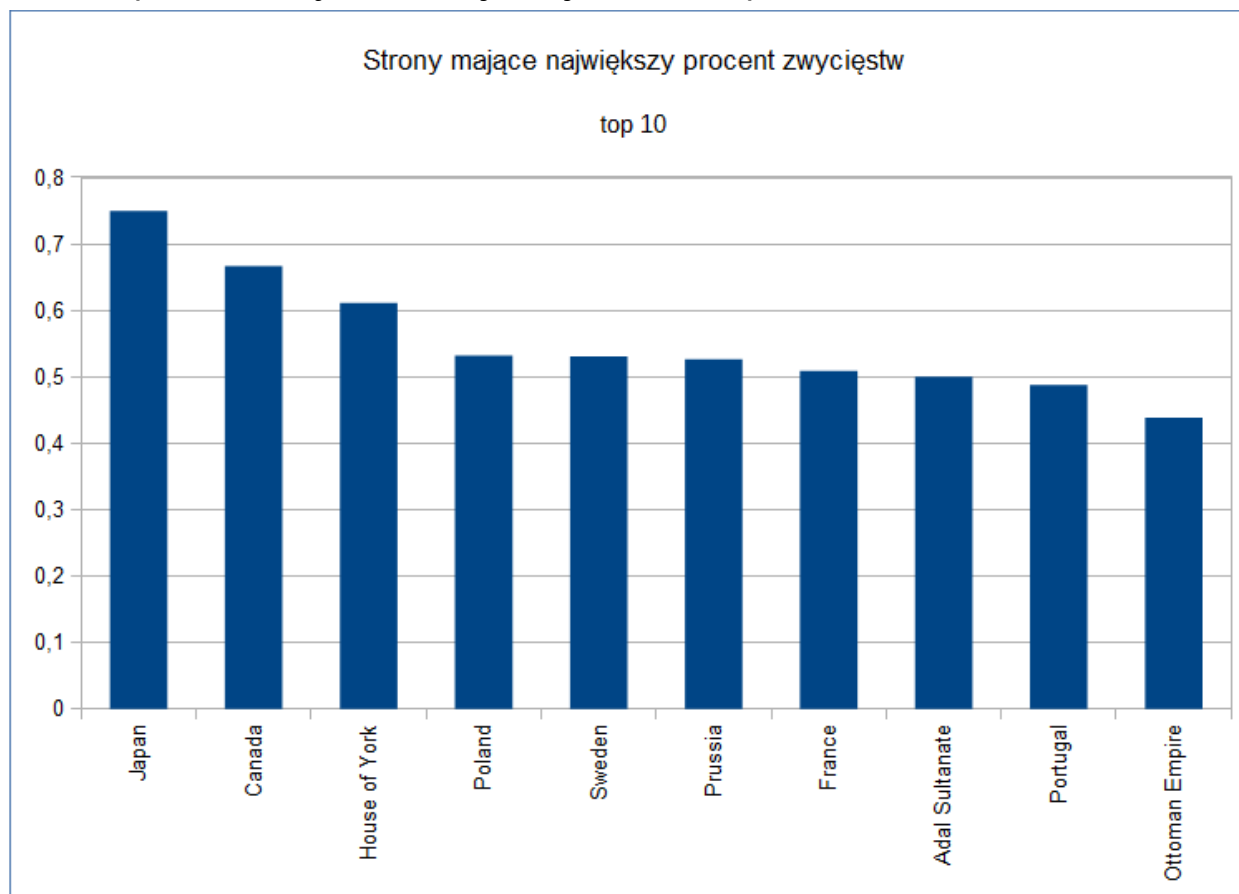
Wykres nr 7: Strony konfliktu z największą liczbą zwycięstw.



Państwo	Liczba zwycięstw
France	113
Spain	46
The Netherlands	34
England	33
Ottoman Empire	32
Sweden	26
Poland	25
Portugal	20
Prussia	20
Great Britain	14
Russia	12
House of York	11
Holy Roman Empire	10
Canada	10
Japan	9
Scotland	8
Hungary	7
Oda Nobunaga	7
Hussite	6
House of Lancaster	6

Dominacja Francji w Europie. Względnie duża liczba zwycięstw Hiszpanii (zapewne konkwistadorzy, Holendrów (kolonie). Co ciekawe, brak Rosji i Austrii w czołówce zestawienia.

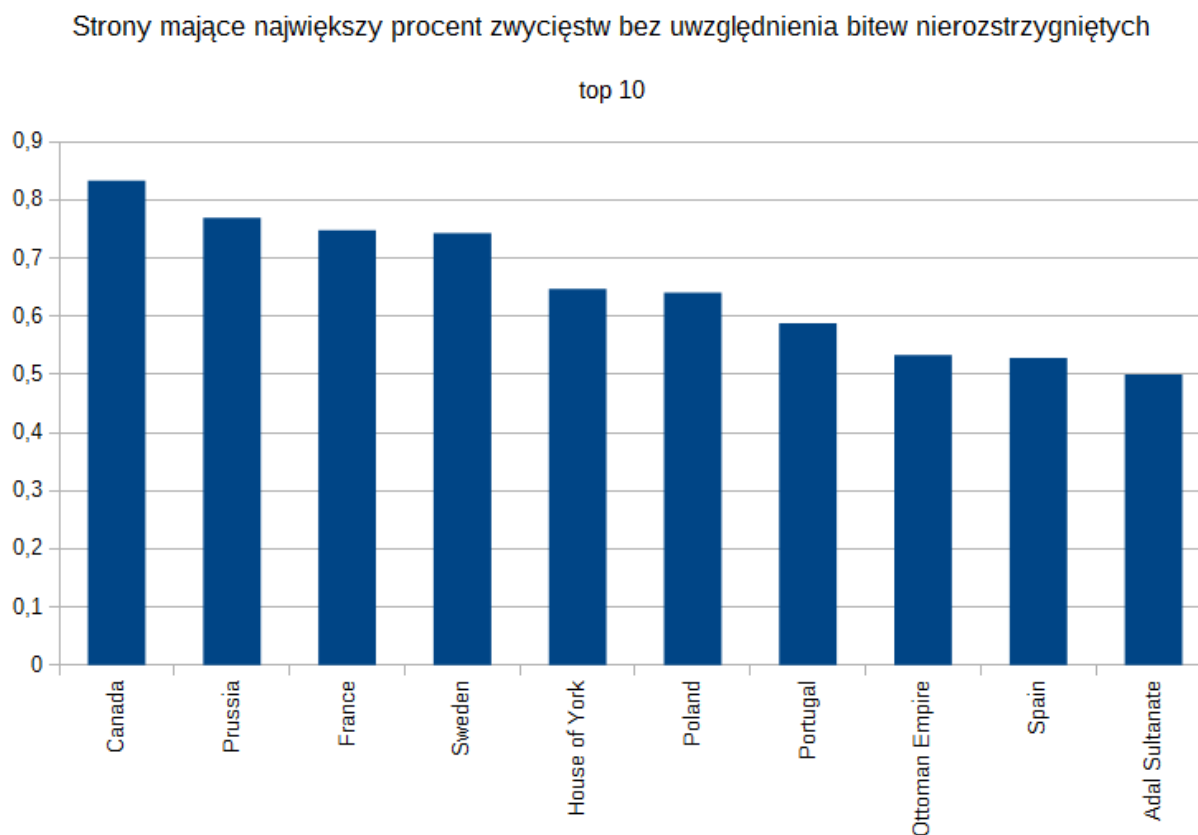
Wykres nr 8: Strony konfliktu z największą procentową ilością zwycięstw w bitwach (uczestniczące w co najmniej 10 bitwach).



Państwo	Procent zwycięstw
Japan	75,00%
Canada	66,67%
House of York	61,11%
Poland	53,19%
Sweden	53,06%
Prussia	52,63%
France	50,90%
Adal Sultanate	50,00%
Portugal	48,78%
Ottoman Empire	43,84%
Takeda Shingen	38,46%
Spain	37,10%
Hungary	36,84%
Moldavia	36,36%
House of Lancaster	33,33%
Serbia	33,33%
Roundhead	33,33%
The Netherlands	32,38%
Oda Nobunaga	30,43%
England	30,00%

W tym zestawieniu najlepszy rezultat uzyskały strony które nie angażowały się w duże konflikty europejskie (Japonia, Kanada). Z europejskiej czołówki brak Anglii/Wielkiej Brytanii.

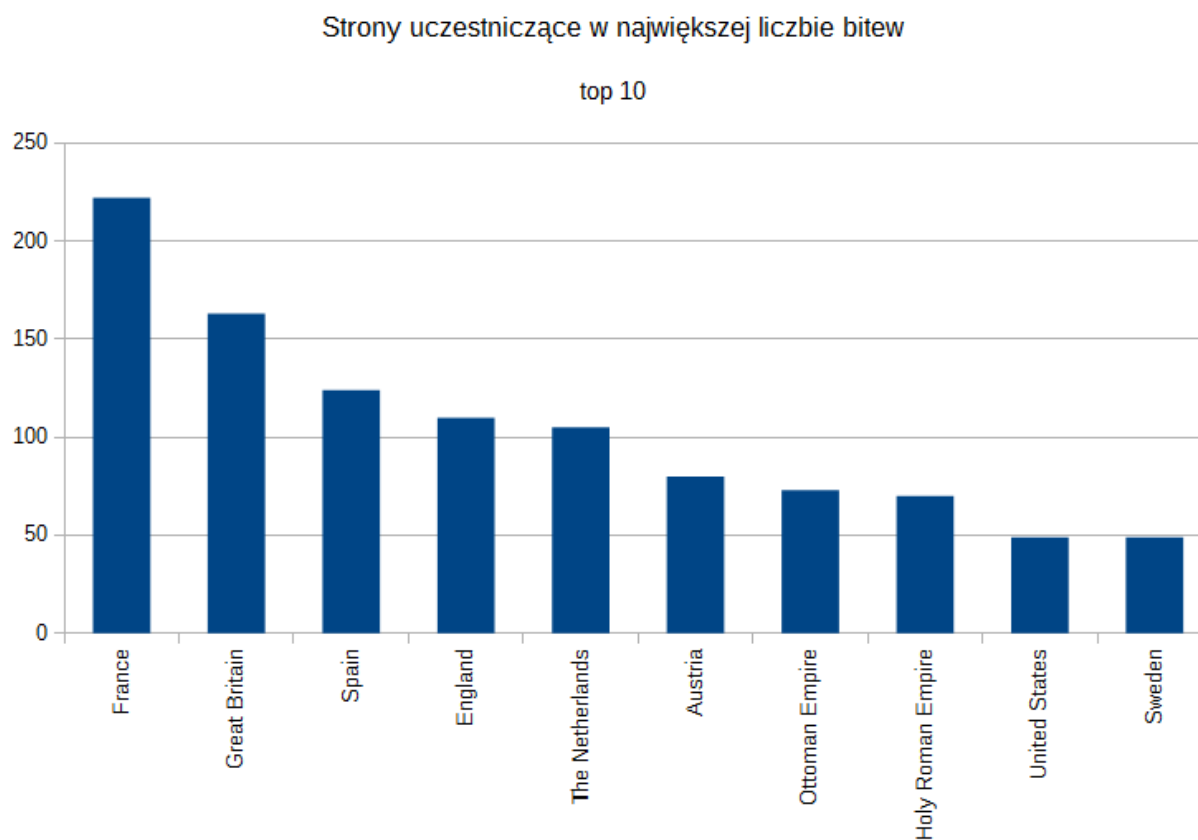
Wykres nr 9: Strony konfliktu z największą procentową ilością zwycięstw w bitwach bez uwzględniania tych nierozstrzygniętych (uczestniczące w co najmniej 10 bitwach).



Państwo	Procent zwycięstw
Canada	83,33%
Prussia	76,92%
France	74,83%
Sweden	74,29%
House of York	64,71%
Poland	64,10%
Portugal	58,82%
Ottoman Empire	53,33%
Spain	52,87%
Adal Sultanate	50,00%
The Netherlands	45,95%
Hungary	43,75%
Oda Nobunaga	43,75%
England	41,77%
Russia	41,38%
House of Lancaster	35,29%
Scotland	34,78%
Great Britain	25,00%
Holy Roman Empire	23,26%
Papal States	20,00%

Wykres podobny do poprzedniego, ale lepiej obrazuje militarną siłę krajów. Wysokie miejsca Prus, Francji i Szwecji które odniosły dużo zwycięstw w tym okresie.

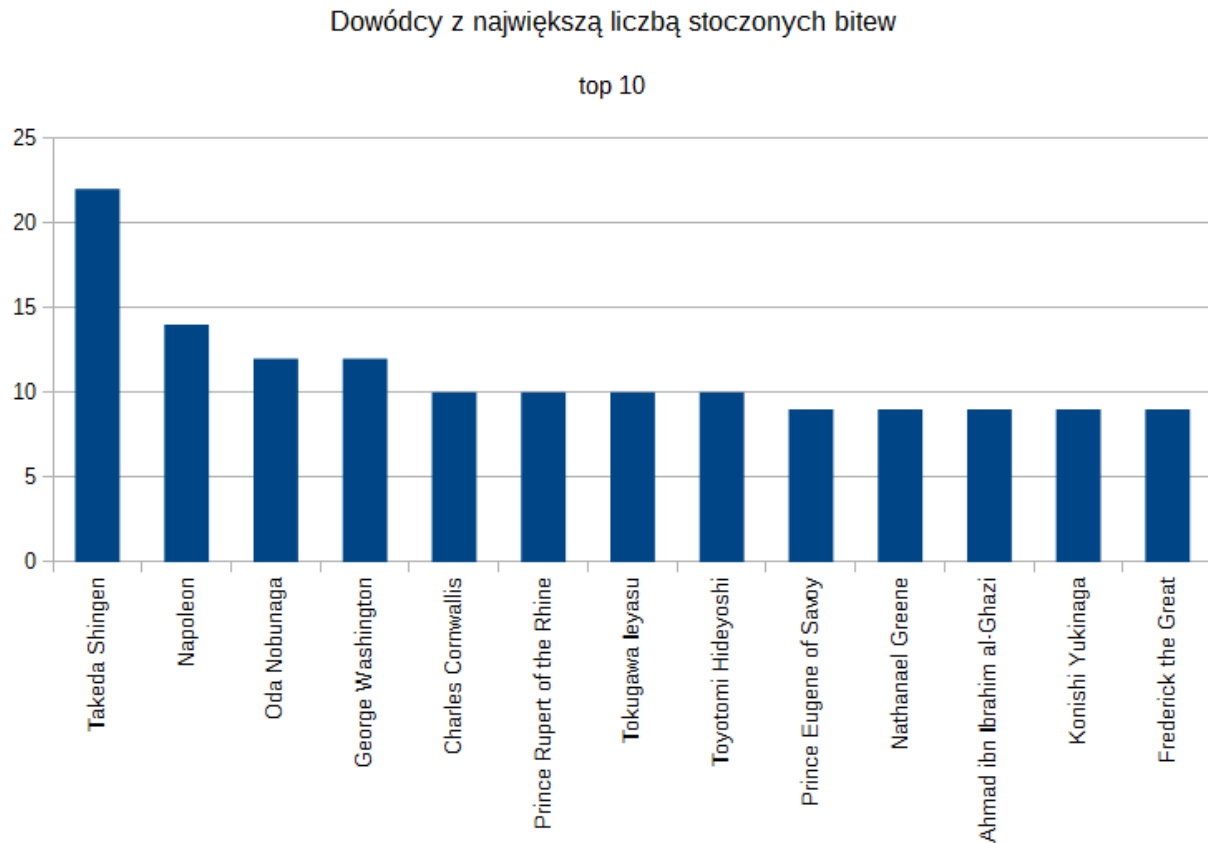
Wykres nr 10: Strony konfliktu z największą ilością stoczonych bitew.



Państwo	Liczba bitew
France	222
Great Britain	163
Spain	124
England	110
The Netherlands	105
Austria	80
Ottoman Empire	73
Holy Roman Empire	70
United States	49
Sweden	49
Poland	47
Russia	44
Portugal	41
Prussia	38
Scotland	33
Oda Nobunaga	23
Republic of Venice	21
Hungary	19
Denmark	19
House of Lancaster	18

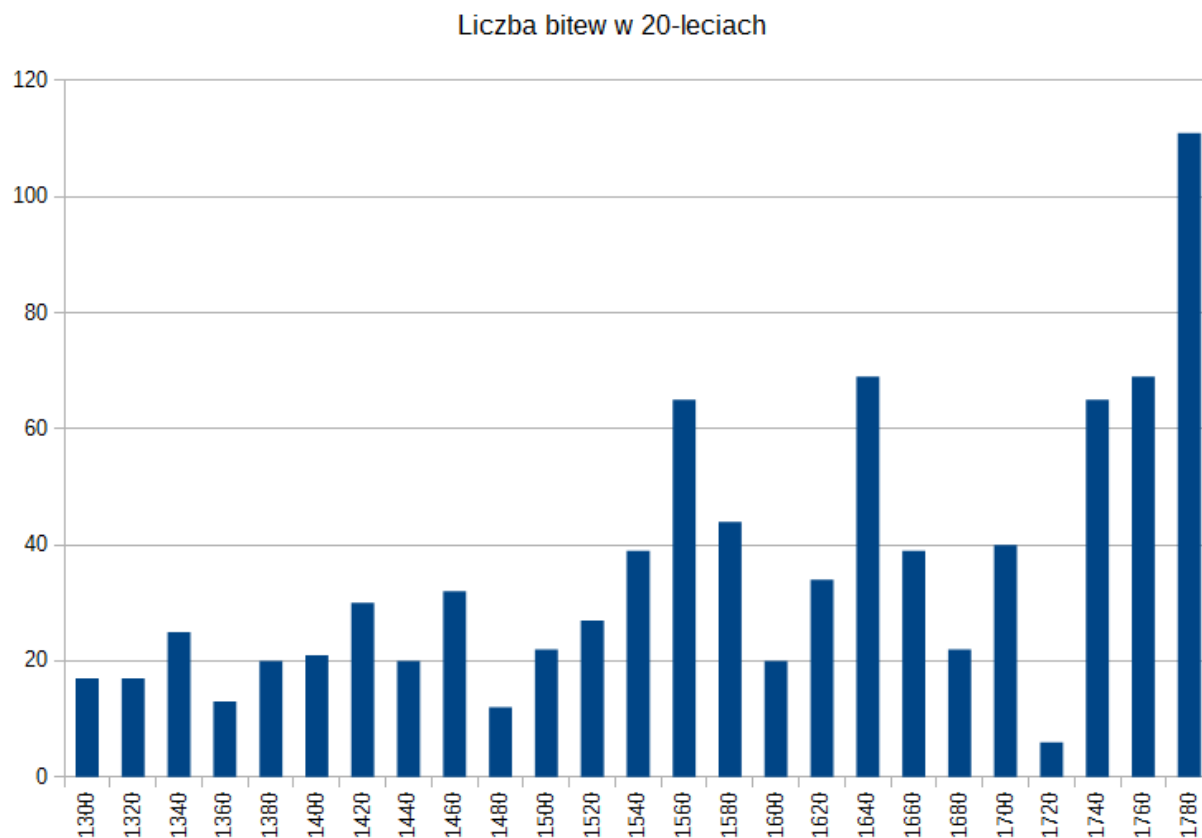
Największe zaangażowanie militarne wykazywały europejskie państwa zachodnie: Francuzi i Anglicy, oraz Hiszpanie i Holendrzy. Ciekawy brak Polski i Rosji.

Wykres nr 11: Liderzy z największą liczbą stoczonych bitew.



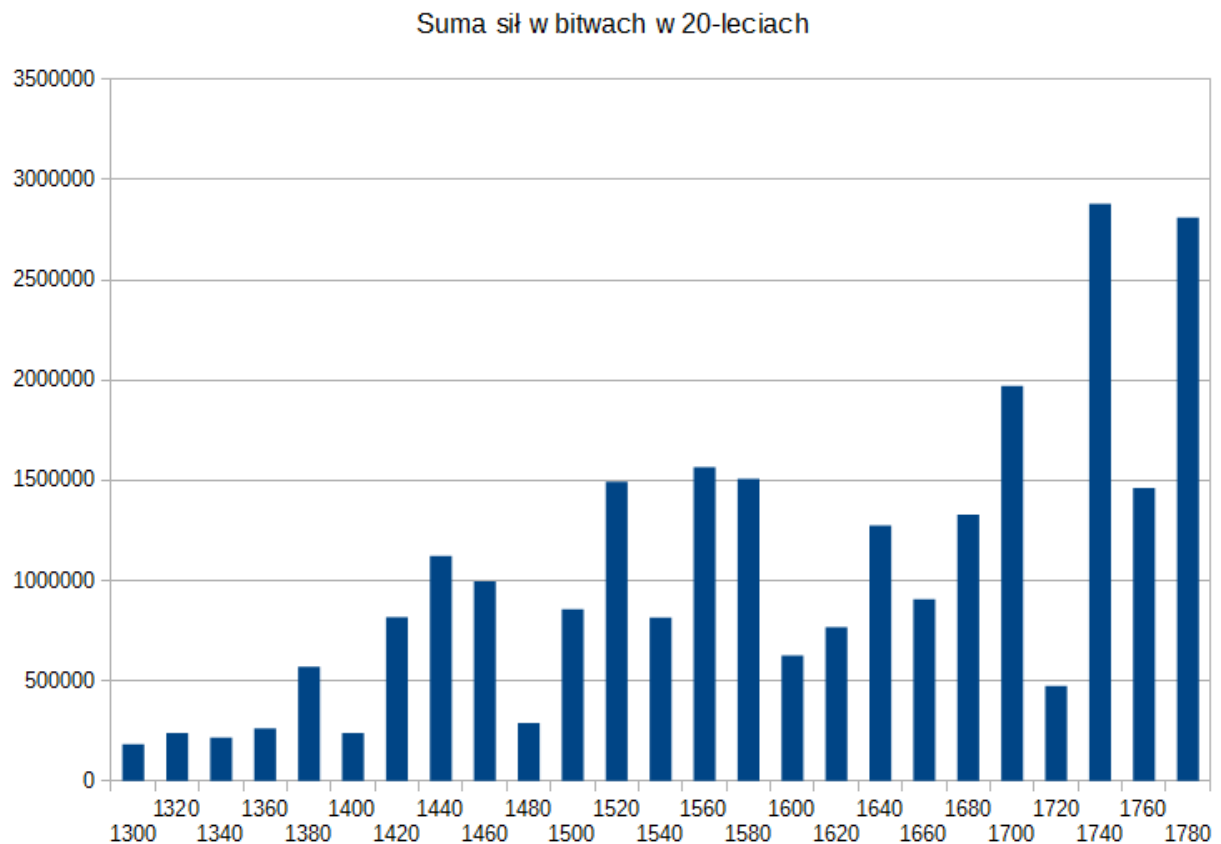
Napoleon - całkiem dobry wynik, mając na uwadze że dane kończą się na roku 1800. W czołówce bardzo dużo dowódców z dwóch okresów: Sengoku Period i amerykańskiej rewolucji, zapewne z powodu dużej ilości bitew w tamtych czasach na małej przestrzeni czasowej.

Wykres nr 12: Liczba bitew na przestrzeni lat (grupowane co 20 lat).



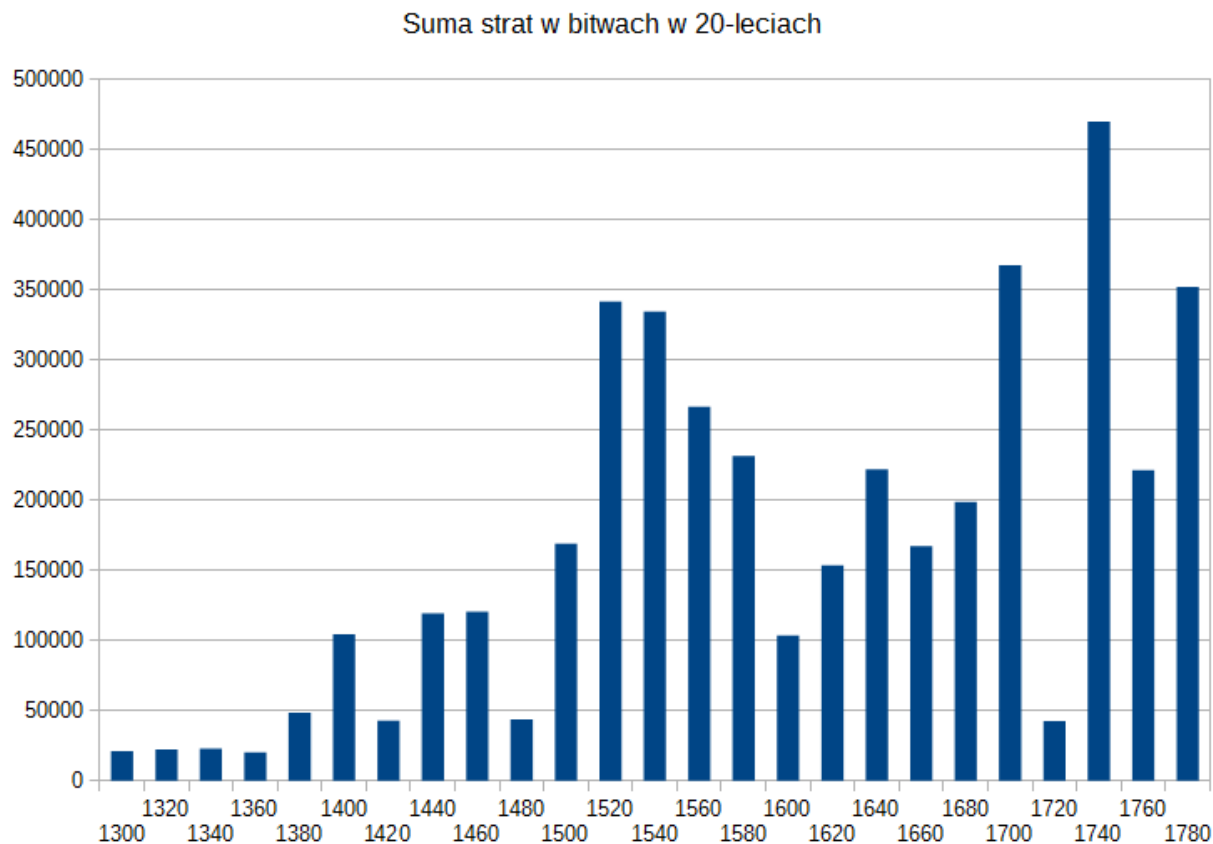
Tendencja mimo wszystko rosnąca, z peakami przypadającymi na duże wojny (trzydziestoletnia, siedmioletnia, amerykańska rewolucja). Zwłaszcza lata 1780-1800 mają bardzo dużą liczbę bitew.

Wykres nr 13: Liczba sił na przestrzeni lat (grupowane co 20 lat).



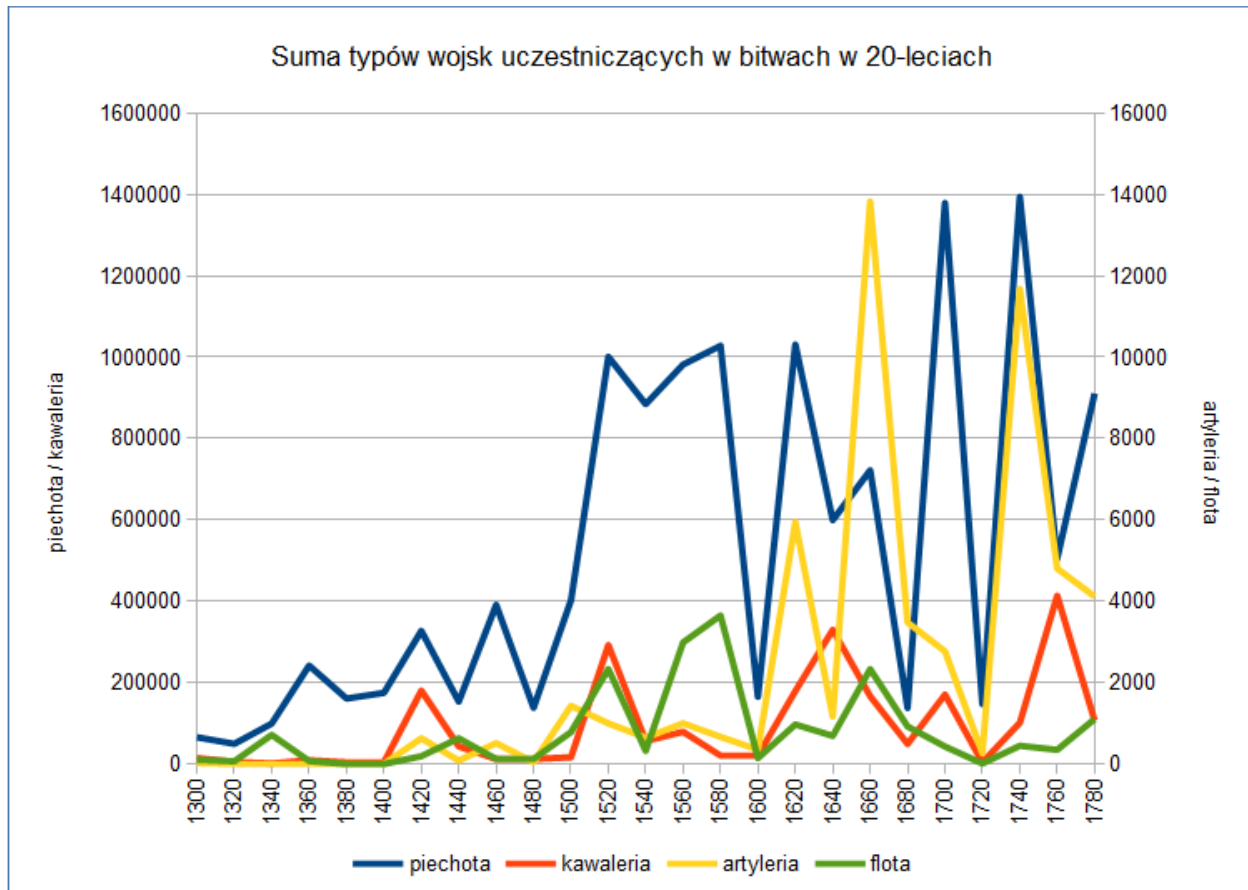
Suma sił pokazuje nieregularny wzrost sił w czasie, jednocześnie sugerując bardzo duże wahania. Może to wskazywać na zmniejszanie ilości konfliktów przy jednoczesnym zwiększaniu ich intensywności.

Wykres nr 14: Liczba strat na przestrzeni lat (grupowane co 20 lat).



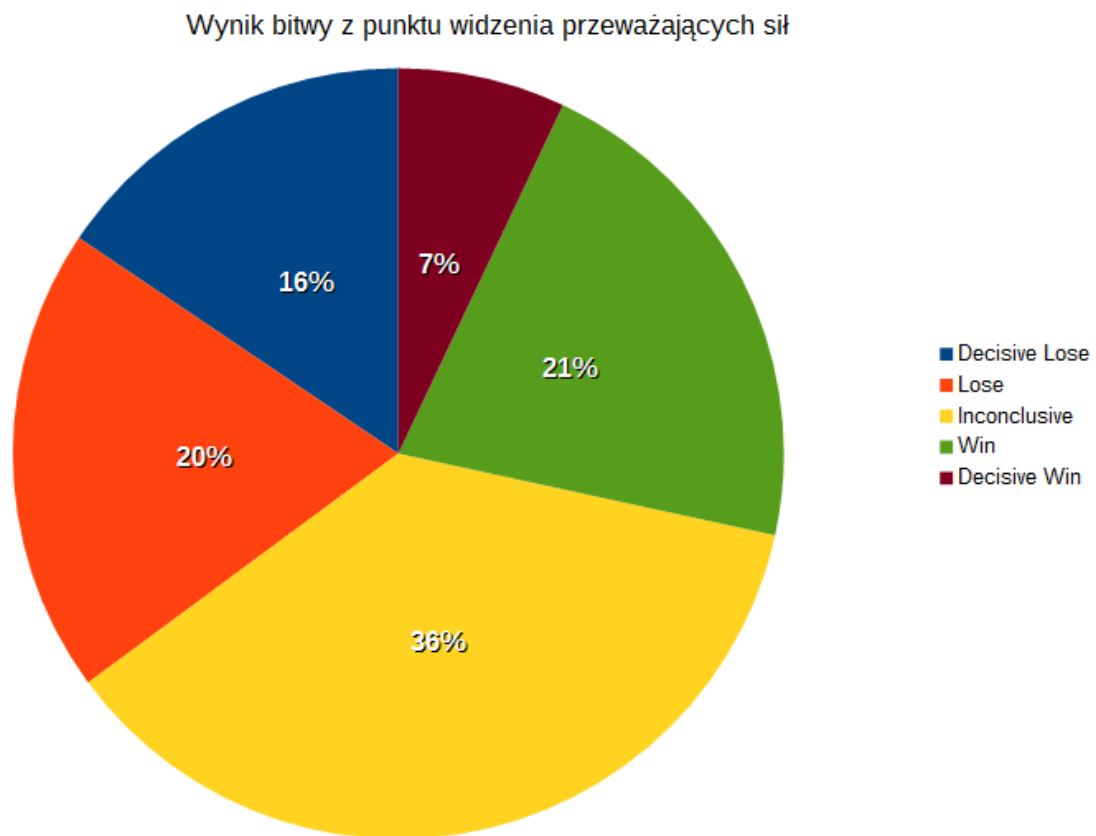
Największe straty przypadły na wieki XVI i XVIII. W tym pierwszym okresie z pewnością duże straty ponosili tubylcy w Nowym Świecie, w tym drugim przyczyną była zapewne globalność konfliktów i zderzenie interesów bardzo wielu państw.

Wykres nr 15: Podział sił na typy jednostek na przestrzeni lat (grupowane co 20 lat).



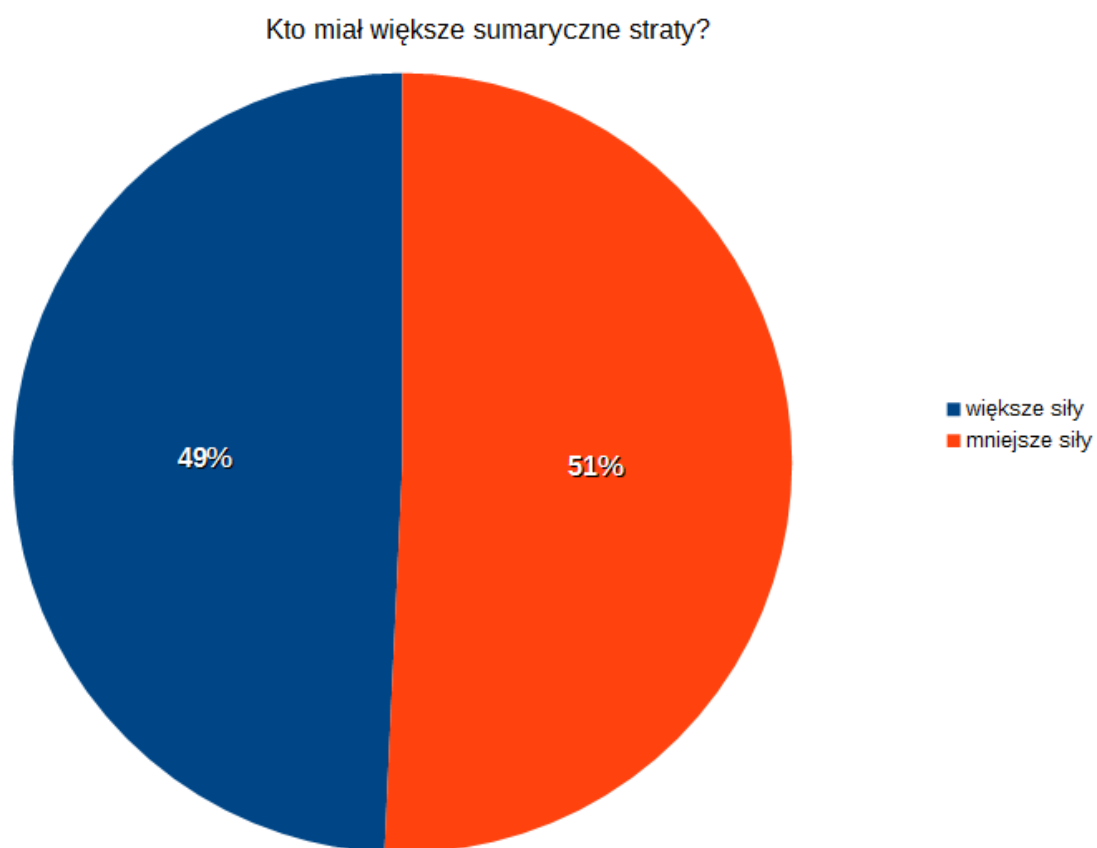
Wykres jest bardzo nieregularny, ale wskazuje na stosunkowo duże użycie piechoty względem kawalerii.

Wykres nr 16: Ile razy zwyciężały strony konfliktu mające przeważające siły w bitwach.



Bardzo ciekawa statystyka, która pokazuje że mając przeważające siły frakcje częściej przegrywały niż wygrywały bitwy. Pewnym rozumowaniem może być to, że mniej liczni wiedzieli o swojej słabości i nie stawali do bitew aż nie mieli dodatkowych argumentów po swojej stronie, których nie da się opisać parametrami ilościowymi.

Wykres nr 17: Ile razy strony konfliktu z przeważającymi siłami ponosiły większe straty.



Ta statystyka okazała się być neutralna - nie widać dominującego związku między ilością strat a wielkością sił.

b. Analizy sieci

Analizowane sieci:

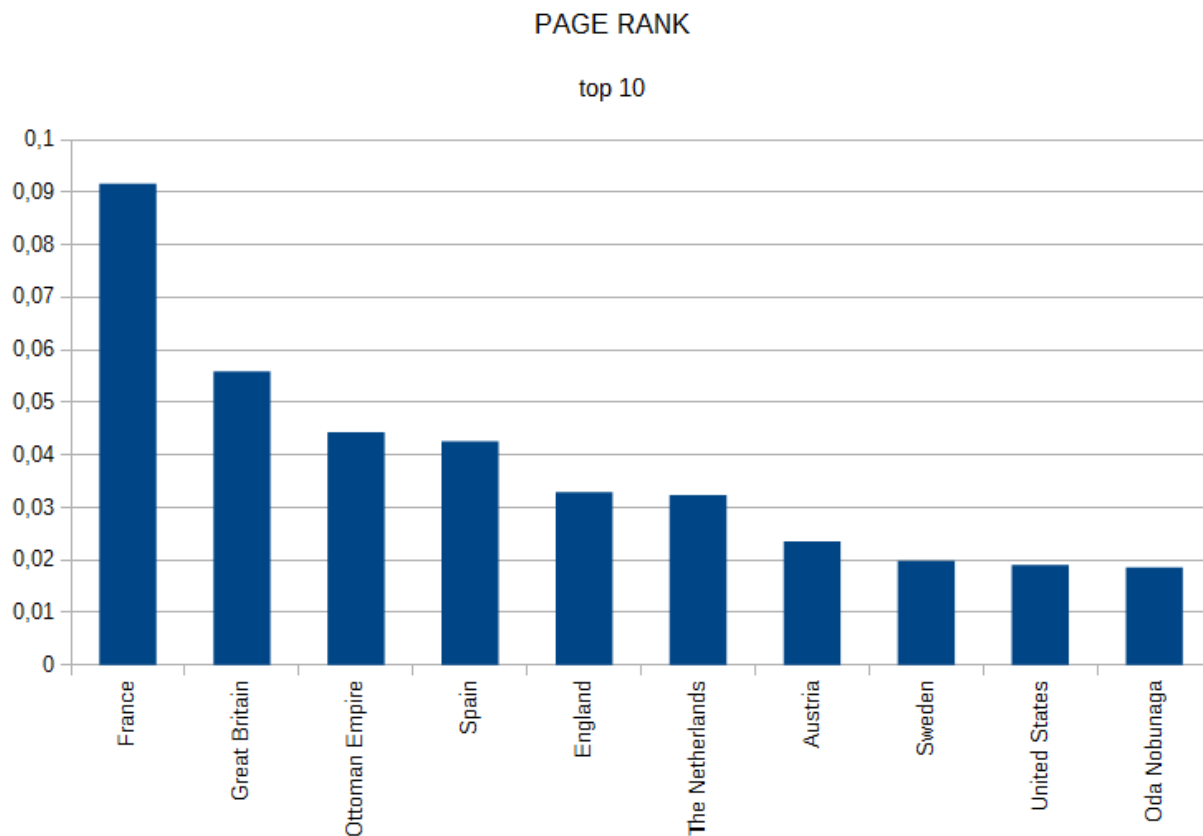
- Sieć nr 1: sieć łącząca strony konfliktu walczące w tych samych bitwach przeciwko sobie (jedna bitwa - jedno połączenie).
- Sieć nr 2: sieć łącząca strony konfliktu walczące w tych samych bitwach po wspólnej stronie (jedna bitwa - jedno połączenie).
- Sieć nr 3: sieć łącząca dowódców walczących w tych samych bitwach, niezależnie po której stronie (jedna bitwa - jedno połączenie).

Podstawowe statystyki sieci:

	sieć 1	sieć 2	sieć 3
średni stopień ważony	15,68	11,88	5,96
średnica	7	8	8
gęstość	0,018	0,032	0,009

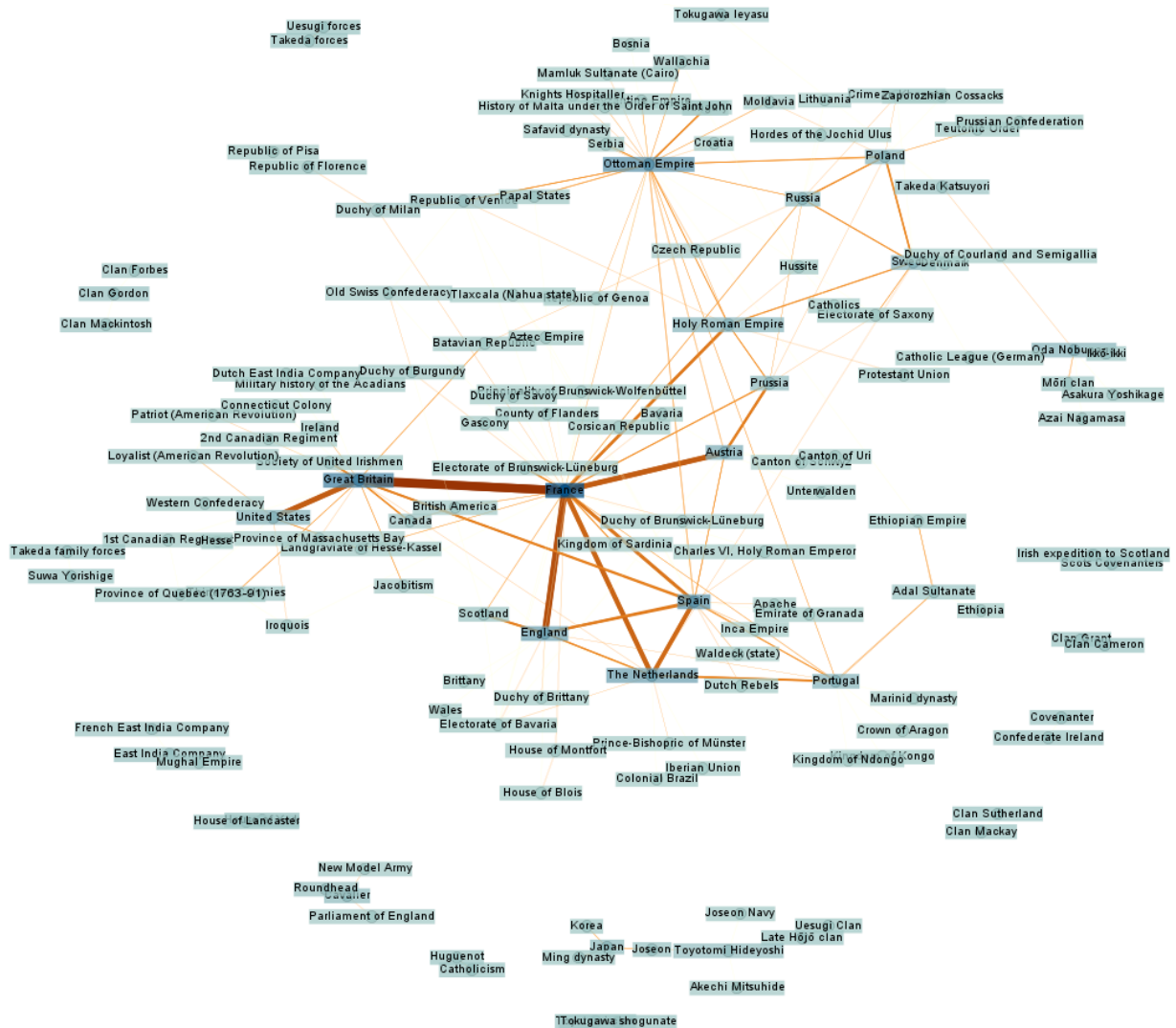
Sieć nr 1: sieć łącząca strony konfliktu walczące w tych samych bitwach przeciwko sobie (jedna bitwa - jedno połączenie):

Strony konfliktu z najwyższym Page Rankiem:



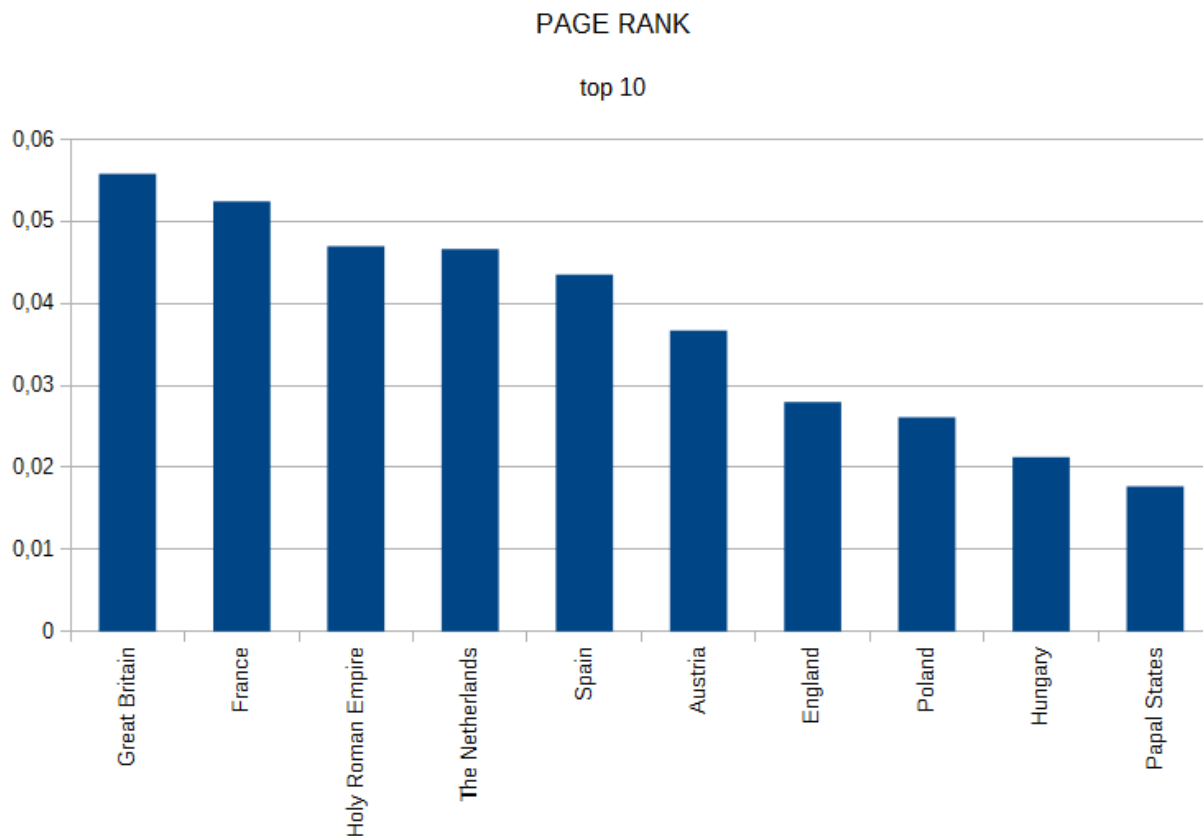
Duży Page Rank Francji i Wielkiej Brytanii nie dziwi. Ciekawa jest obecność na trzecim miejscu Imperium Osmańskiego, które nie było w czołówce liczebności bitew. Jednak w tym konkretnym przypadku decyduje wysoka 'jakość' przeciwników z którymi walczyli.

Wizualizacja sieci:



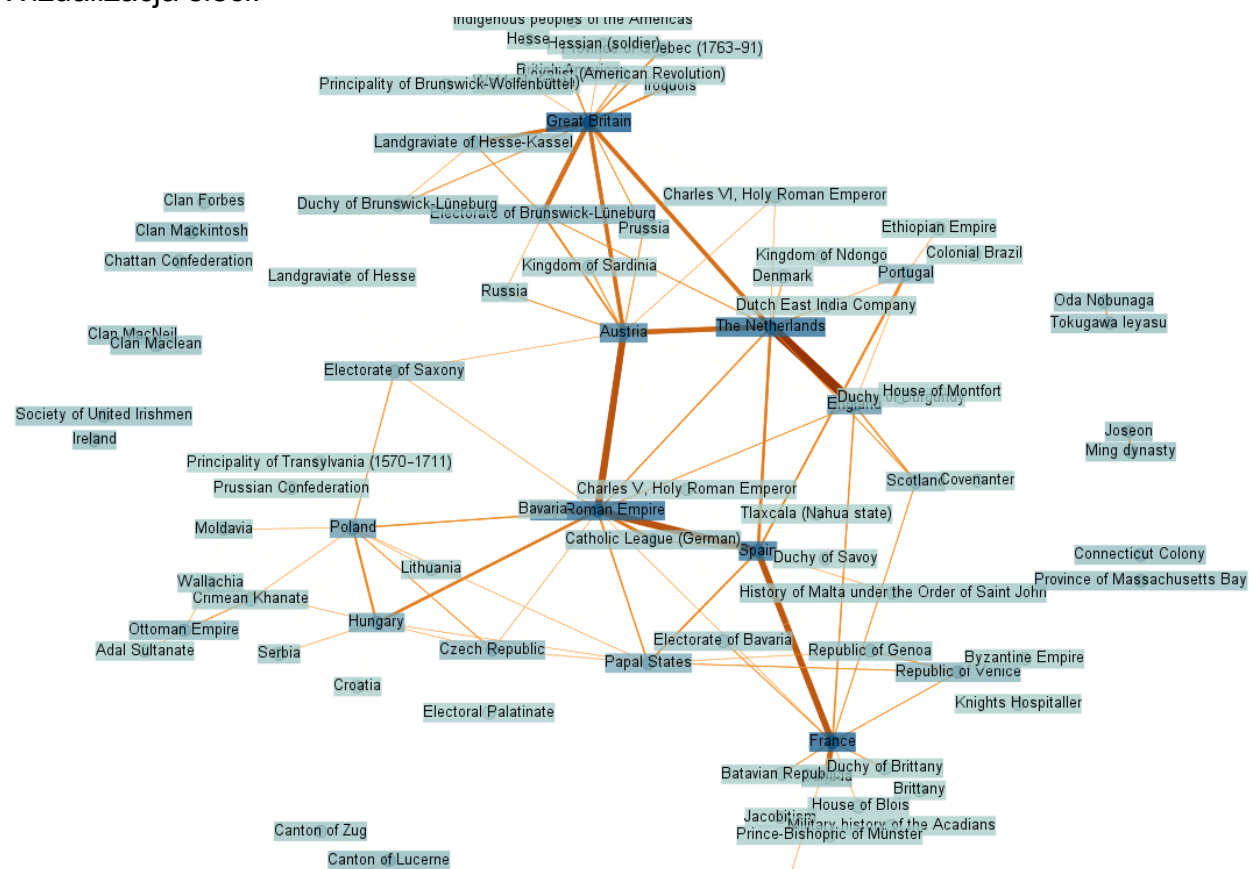
Sieć nr 2: sieć łącząca strony konfliktu walczące w tych samych bitwach po wspólnej stronie (jedna bitwa - jedno połączenie):

Strony konfliktu z najwyższym Page Rankiem:



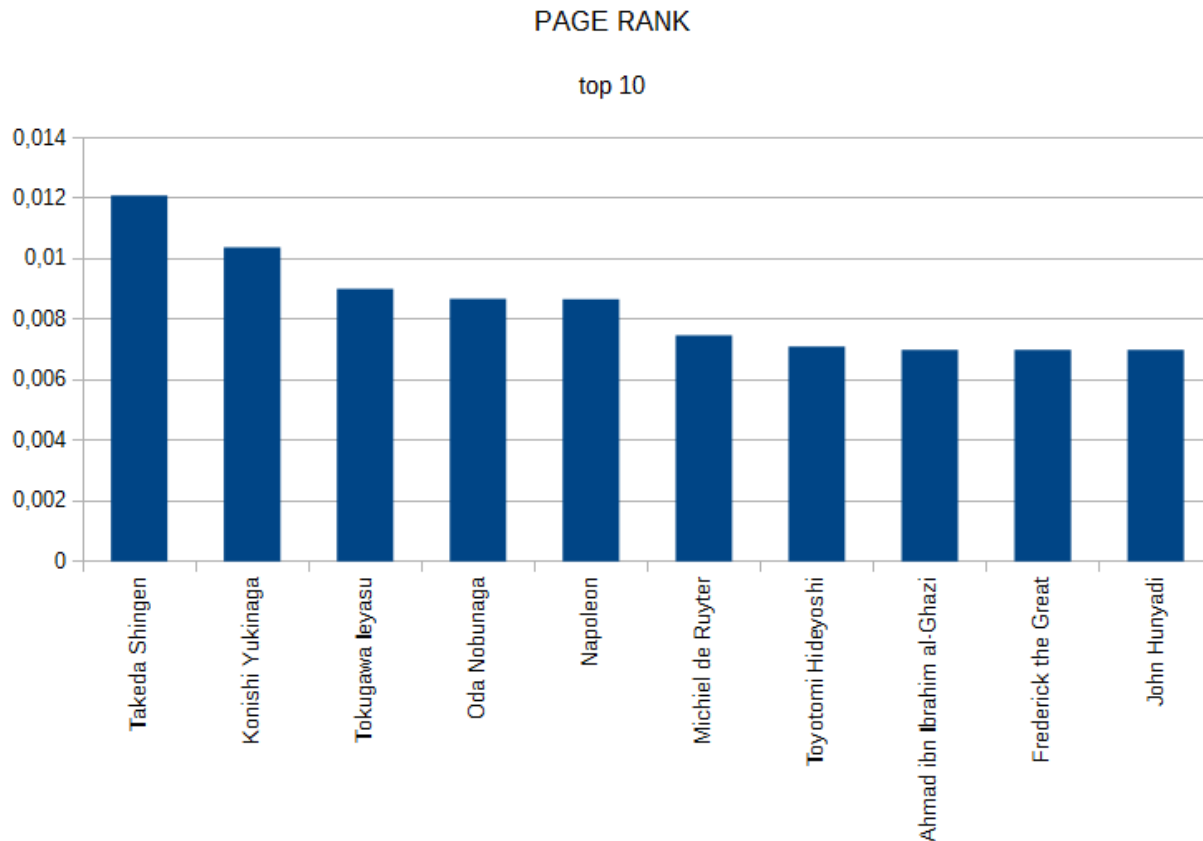
W tej sieci widoczna jest raczej jakość dyplomacji państw. W porównaniu z poprzednią siecią, znikły Imperium Osmańskie i Szwecja które najczęściej walczyły same przeciwko innym. Obecność Świętego Cesarstwa Rzymskiego i Papiestwa w czołówce wskazuje na duży wpływ tych ośrodków władzy i jednocześnie przyczynę wielu z analizowanych bitew.

Wizualizacja sieci:



Sieć nr 3: sieć łącząca dowódców walczących w tych samych bitwach, niezależnie po której stronie (jedna bitwa - jedno połączenie):

Dowódcy z najwyższym Page Rankiem:



Jak widać, czołówka została zdominowana przez Japońskich przywódców z czasów Sengoku. Poza nimi są też jednak bardzo znani dowódcy: Napoleon i Fryderyk Wielki.



11. Podsumowanie

Względem oryginalnych planów zrealizowaliśmy cały harmonogram. Udało się zarówno sparsować dane z wikipedii (bitwy, wojny, pokoje) jak i umieścić je w bazie danych w odpowiednich strukturach. Przy przetwarzaniu nie udało się otrzymać niektórych informacji o pokojach (m.in. związki pokojów z wojnami/bitwami) gdyż sposób przechowywania tych informacji na wikipedii nie pozwolił na to. Z tego powodu zrezygnowaliśmy z analiz dotyczących pokojów i skupiliśmy się na bitwach i wojnach.

Jeśli chodzi o przyszły rozwój, można poszukać grup w społecznościach przy pomocy narzędzia CFinder. Można także znaleźć inne źródło niż angielska wikipedia, pobrać pokoje i połączyć relacjami z bitwami i wojnami. Inną opcją jest też rozważenie innego przedziału czasowego.

Lista źródeł

1. <http://en.wikipedia.org>