Przetwarzanie dużych zbiorów danych

Krótki opis usługi

Usługa umożliwia realizację zadań obliczeniowych dla bardzo dużych zbiorach danych lub też wielu bardzo małych zadań w jak najkrótszym czasie. Budowane środowisko umożliwia złożone analizy i obliczenia w takich problemach jak:

- analiza wydźwięku dla danych tekstowych,
- przetwarzanie języka naturalnego,
- analityka marketingowa i analiza danych konsumenckich,
- przetwarzanie sieci złożonych, w tym np. sieci społeczne, sieci ekonomiczne, sieci transportowe,
- uczenie maszynowe i wspomaganie decyzji.

W ramach usługi można także dokonywać wizualizacji otrzymanych wyników badań. Przykładem takich wizualizacji mogą być grafiki prezentujące słowa nacechowane pozytywnie czy negatywnie (rysunki poniżej, im większy rozmiar czcionki danego słowa tym bardziej pozytywne/negatywne jest dane słowo). Na rysunkach przedstawiony jest wynik budowania słownika wykorzystywanego podczas analizy wydźwięku dla języka naturalnego.

Chmura słów - pozytywne zwroty



Chmura słów - negatywne zwroty



Ograniczenia w korzystaniu

Proszę sprawdzić czy aplikowano o wszystkie niezbędne usługi w Portalu PLGRID. Więcej wymaganiach Complex Networks zamieszczono w osobnym rozdziale.

Pierwsze kroki

Zlecanie zadań obliczeniowych z użyciem QCG

Należy wybrać jedną z opcji: **Zleć zadanie QCG (formularz do zlecania) albo Edytor QCG (możliwość stworzenia własnego skryptu Pyton/bash do wykonania obliczeń)**

Przykładowy skrypt QCG dostępny tutaj: **qcg-test-bash.qcg**

Dodatkowy plik tekstowy, dla którego będzie wykonane zliczenie liczby różnych słów w nim występujących: **bible-kjv.txt**



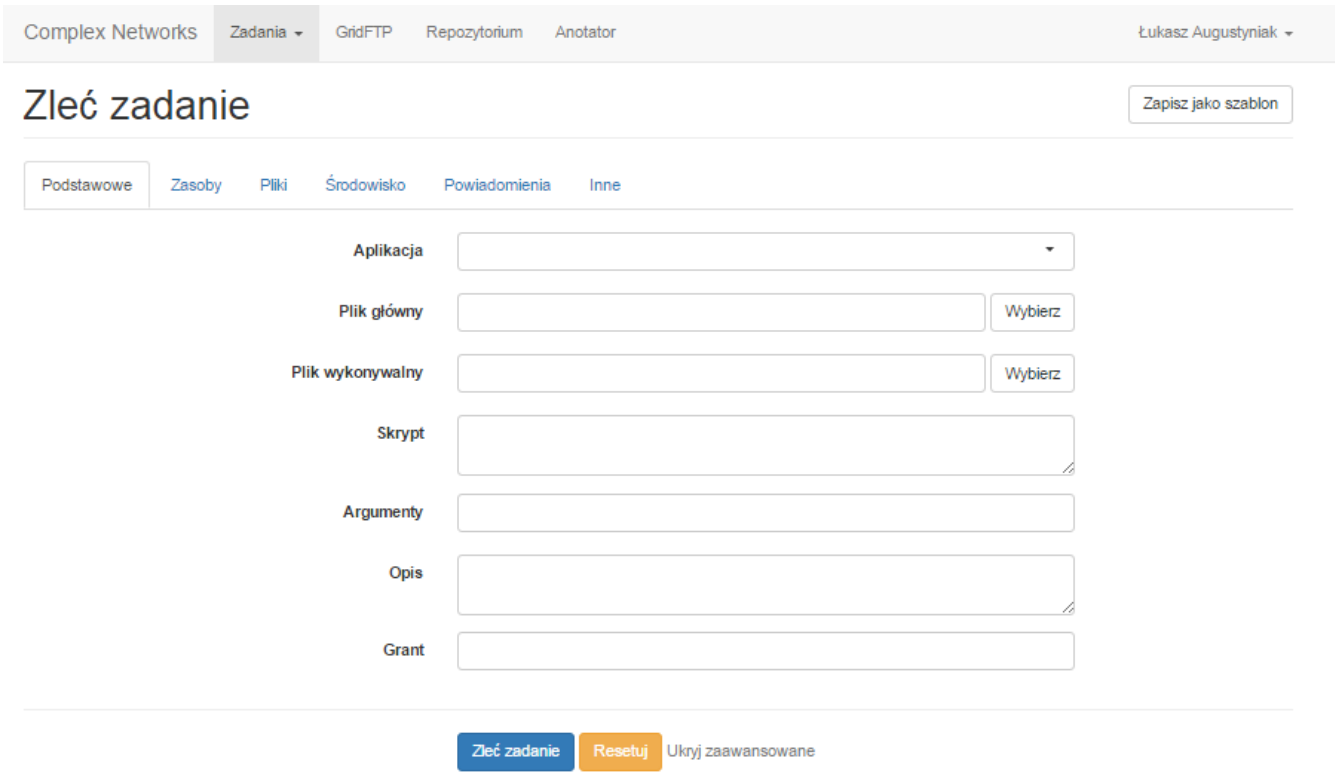
Zleć zadanie QCG

Complex Networks umożliwia zlecanie zadań obliczeniowych z wykorzystaniem poniżej przedstawionego formularza. Wszystkie parametry oraz ich znaczenie można znaleźć na stronie Format opisu zadań.

Po wskazaniu parametrów zadania możemy je uruchomić klikając na przycisk **Zleć zadanie.**

Każde z uruchomionych przez Portal Complex Networks zadań można podejrzeć w zakładce **Zadania -> Moje zadania.**

Ważnym punktem jest możliwość zapisuj specyfikacji zadania w postaci Szablonu - **Zapisz jako szablon**. Dzięki temu możemy zlecać podobne albo wręcz takie same zadania szybciej i prościej.



Przykładowe zadanie zlecane przez formularz - zliczanie słów dla zadanego pliku

Należy przygotować dwa pliki - skrypt bash oraz tekstowy plik do zliczenia słów (nazwa pliku jest bardzo ważna w tym przypadku, ponieważ w skrypcie bash jest zapisana nazwa pliku przykładowego)

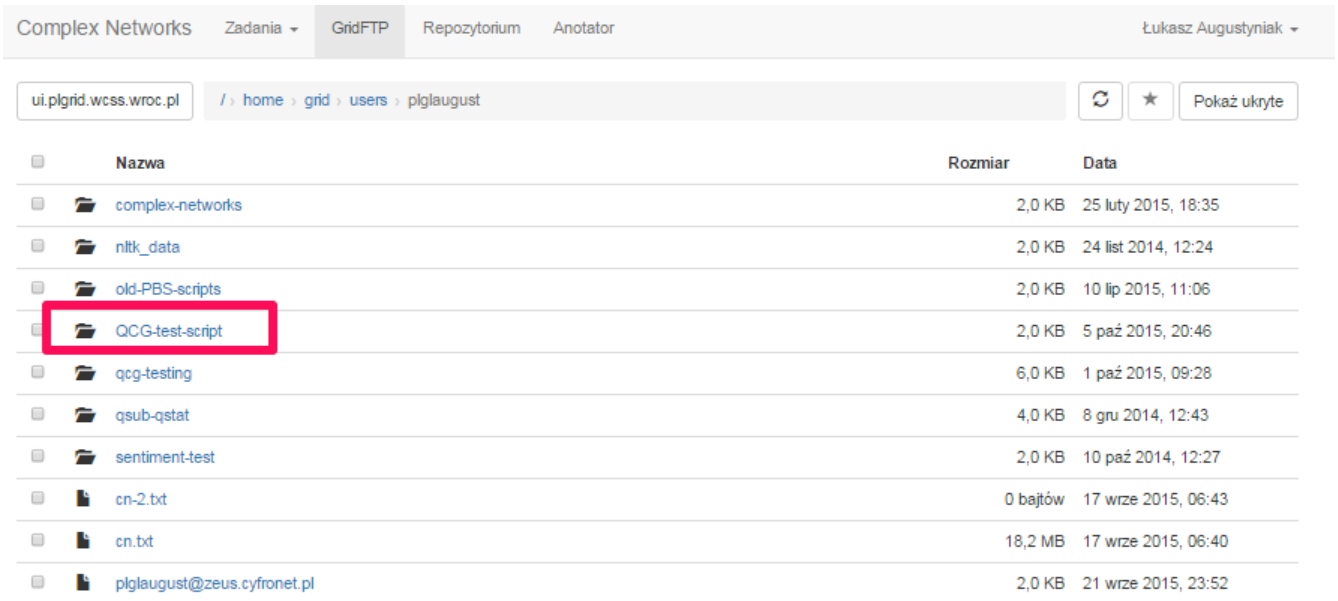
qcg-test-bash.qcg

bible-kjv.txt

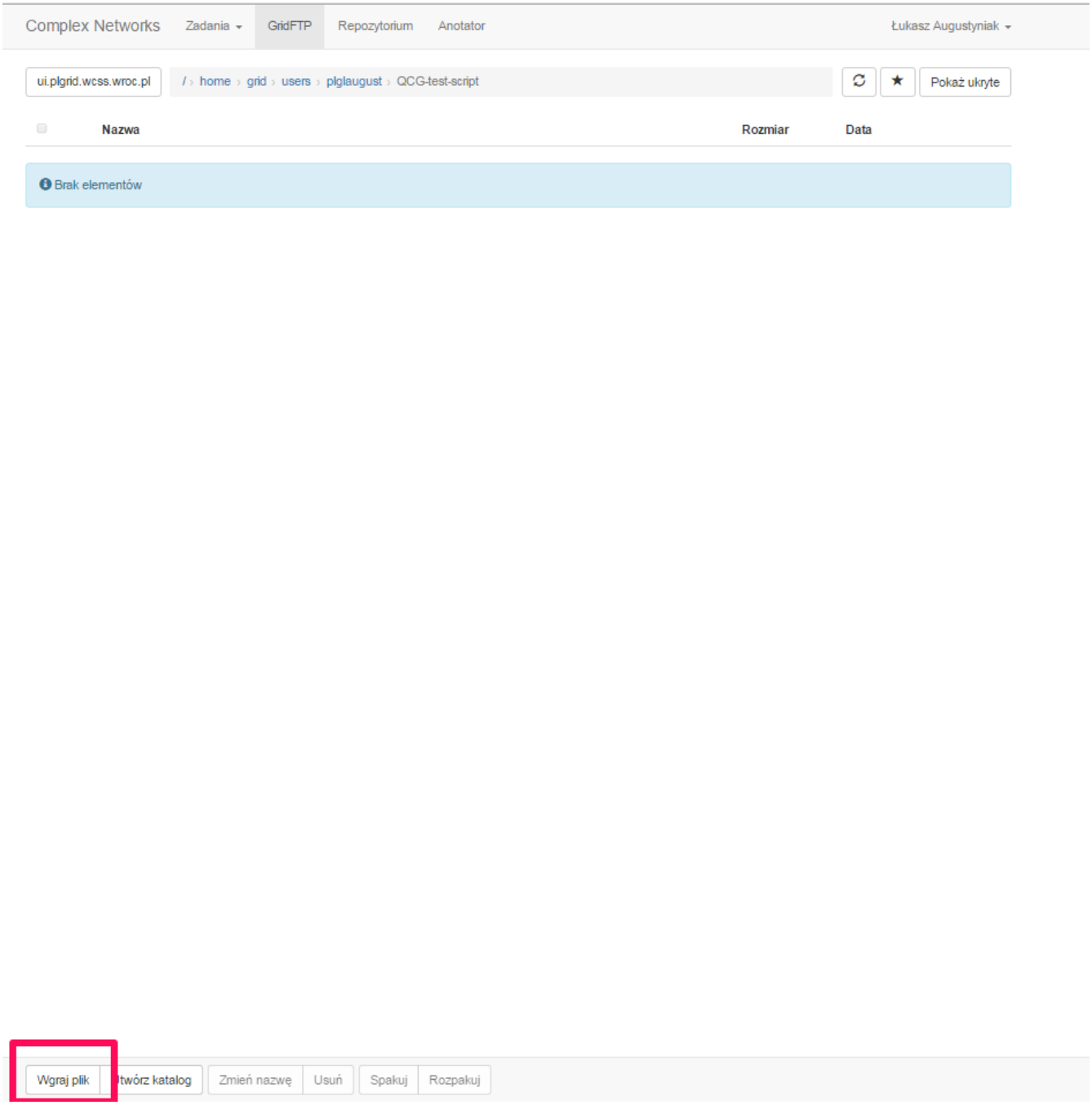
Najpierw musimy dodać pliki na jedną z naszych maszyn dostępowych - możemy połączyć się do nich poprzez SSH Dostęp do UI (User Interface) alb o wykorzystanie GridFTP w Portalu Complex Networks.



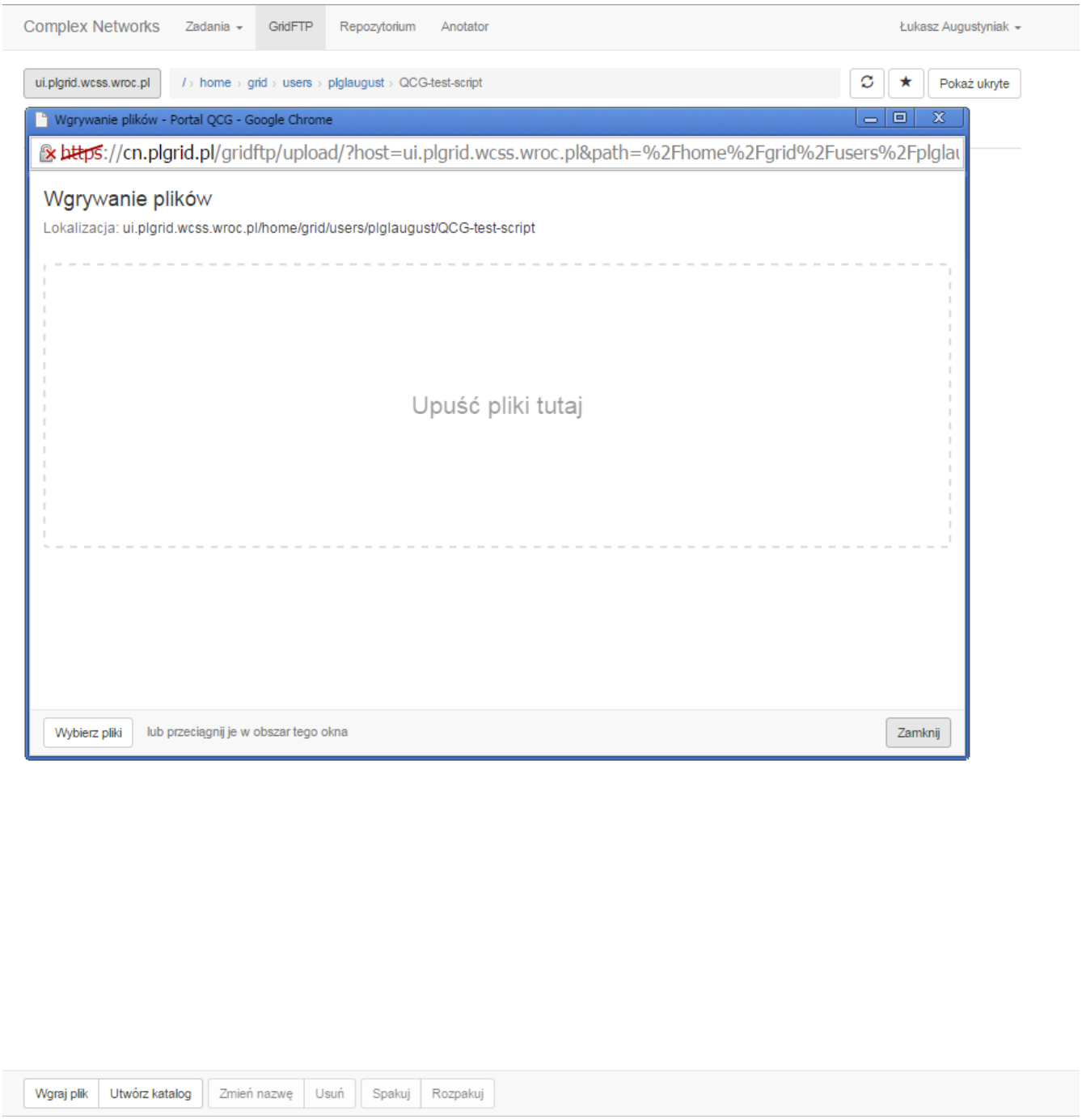
Następnie tworzymy folder `QCG-test-script` i przechodzimy do niego.



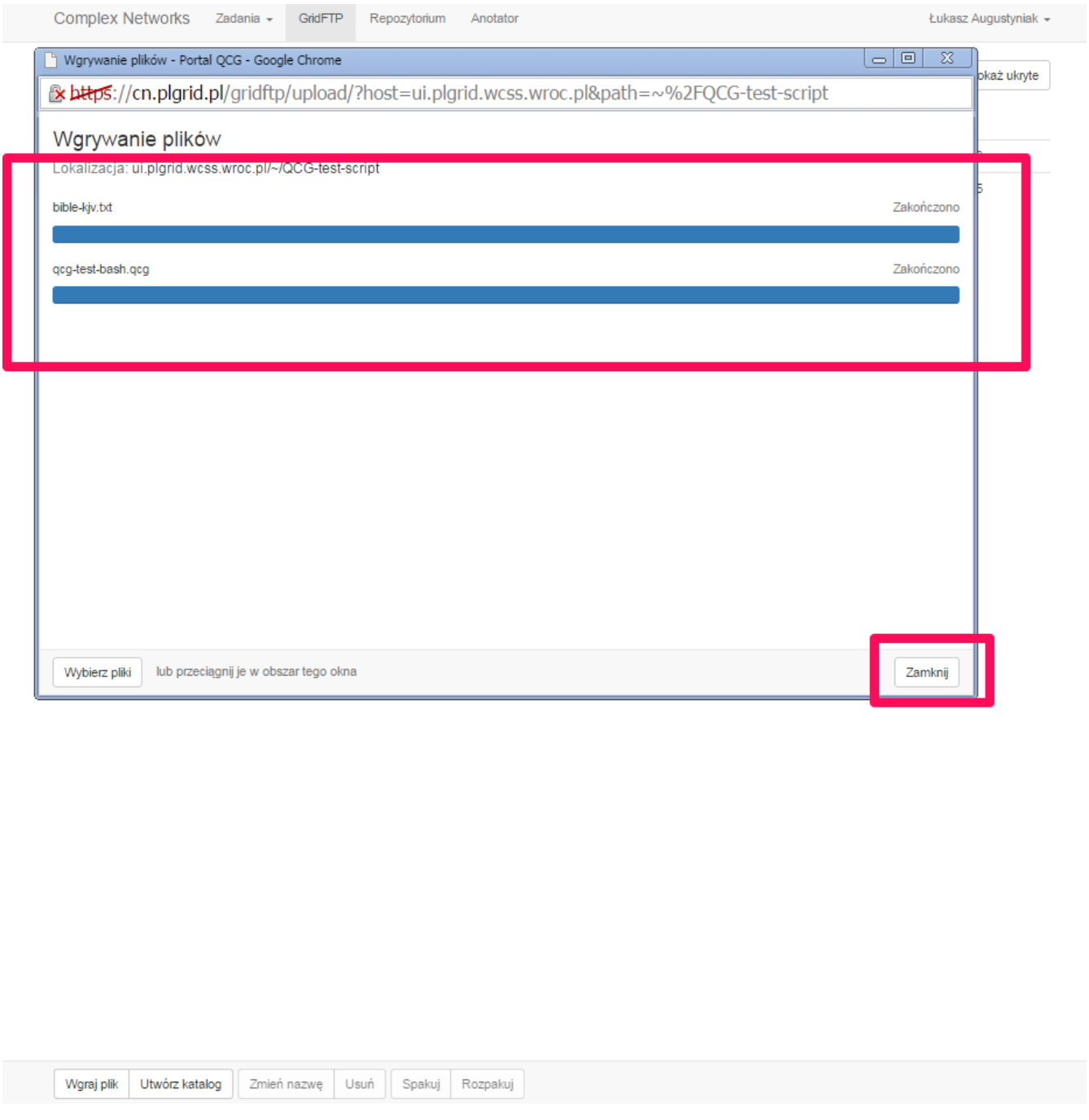
Do niego dodajemy pobrane wcześniej pliki - klikamy w przycisk **Wgraj.**



Pojawi się nowe okienko, do którego metodą przeciągnij/upuść albo poprzez kliknięcie w przycisk **Wybierz plik** dodajemy pliki potrzebne do uruchomienia zadania.



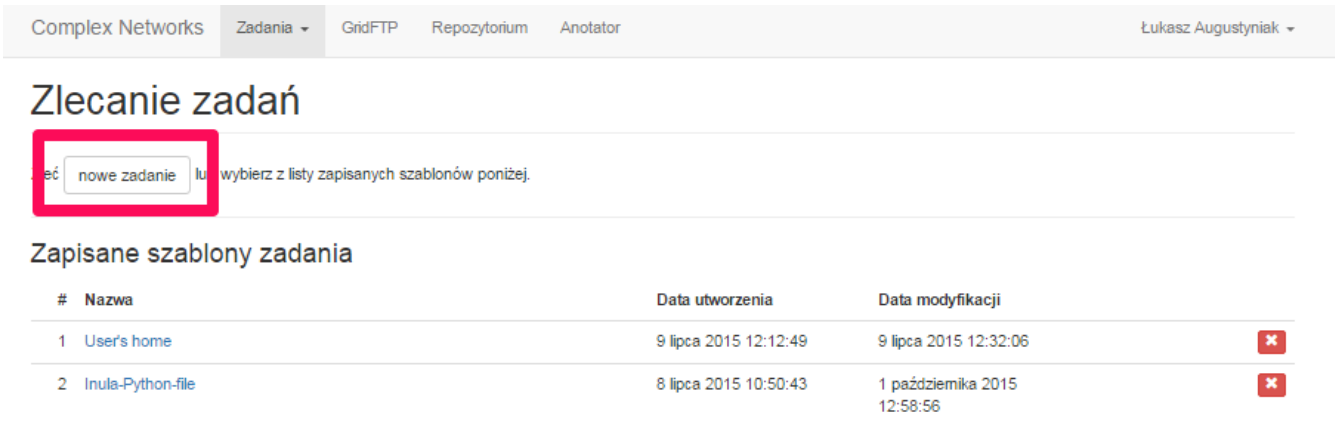
Poniżej widzimy poprawnie wgrane dwa pliki.



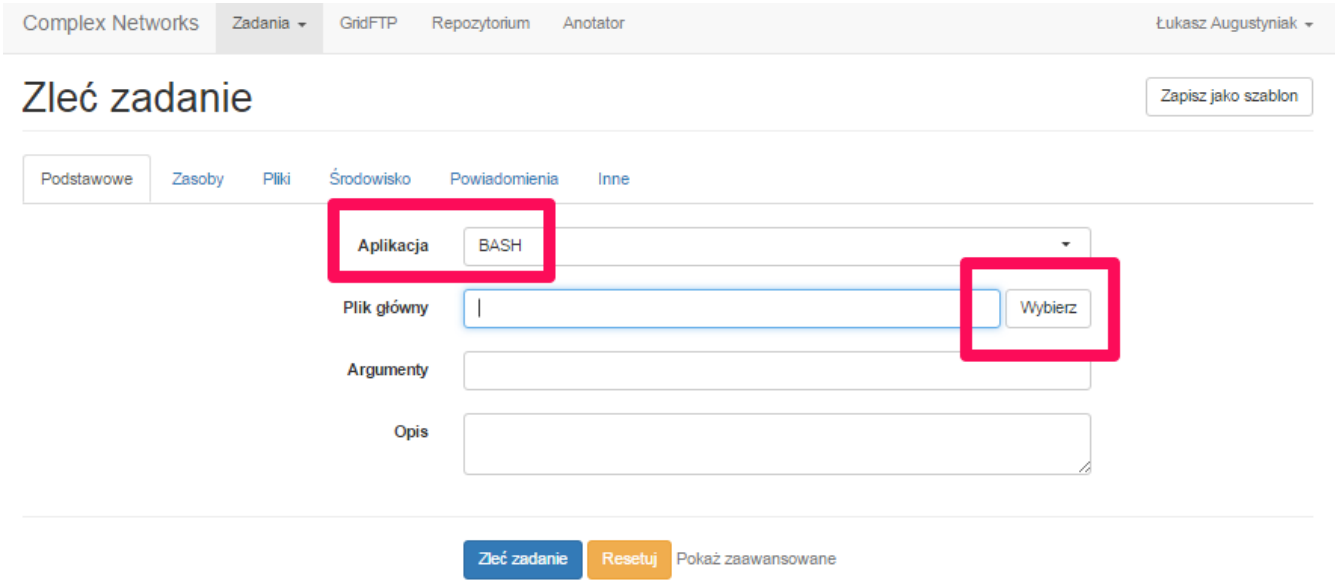
Następnie przechodzimy do formularza **zlecania zadań QCG**.



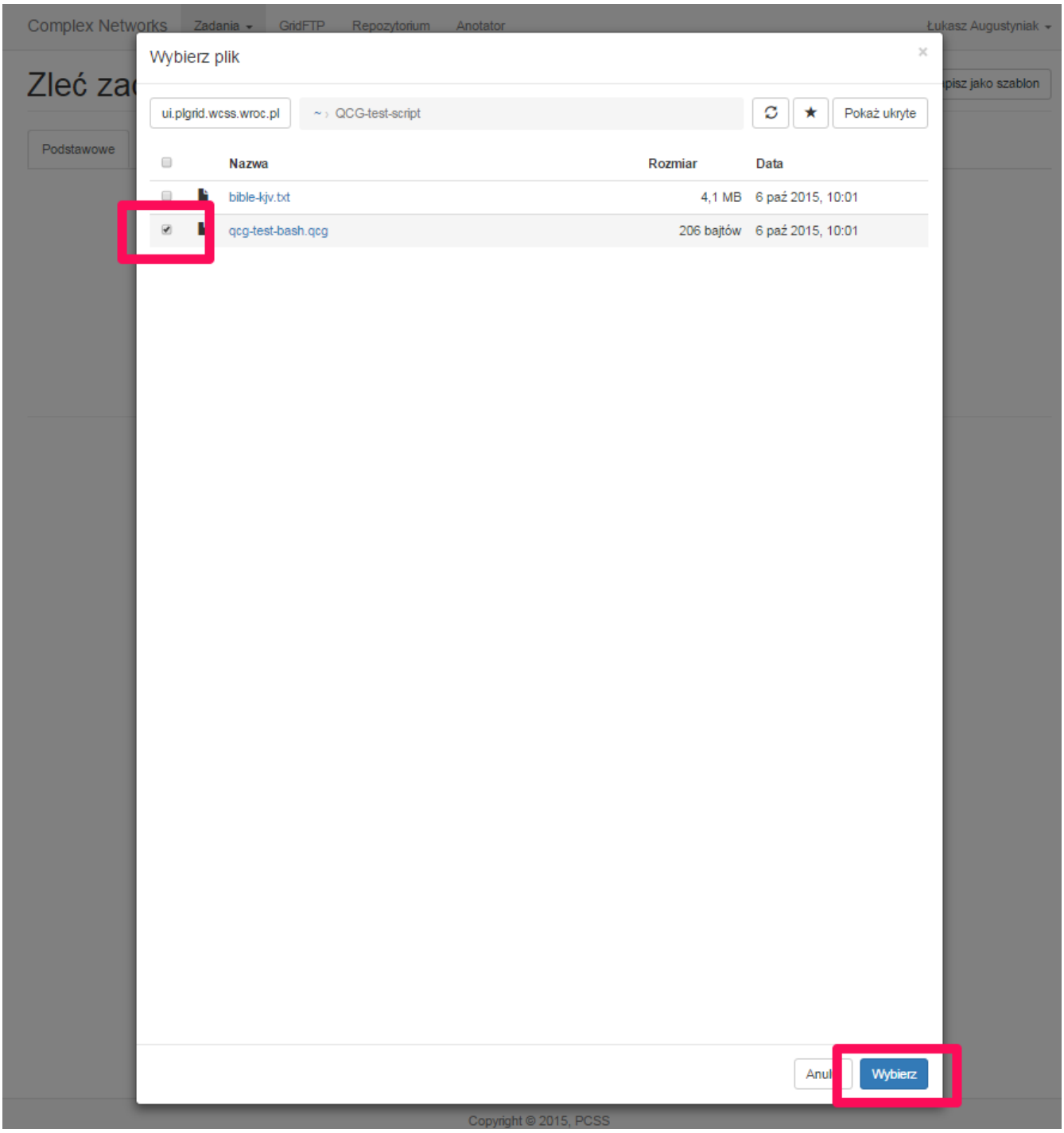
Wybieramy **Nowe zadanie**



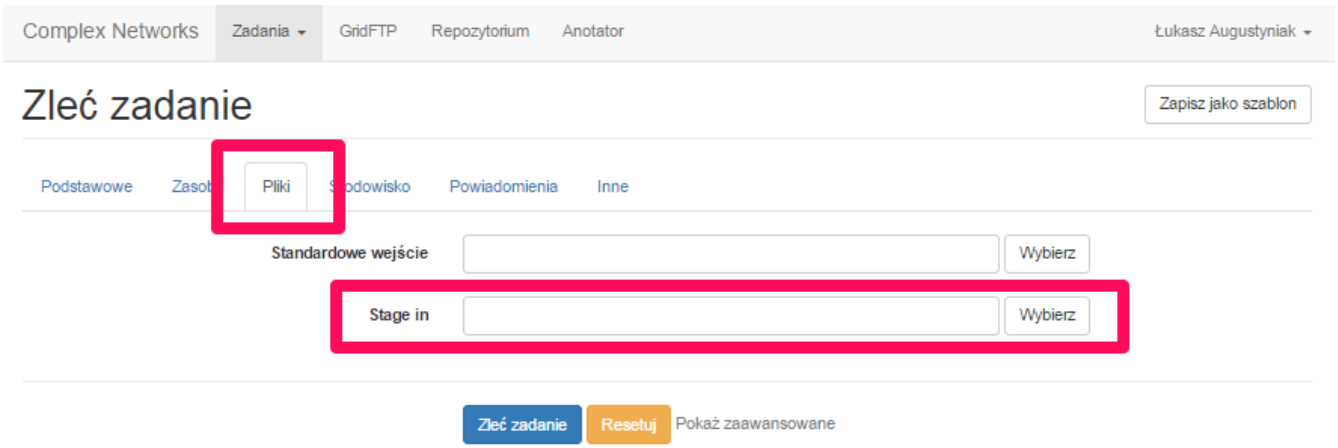
Wybieramy aplikację **BASH** (skypt bash zostanie użyty do uruchomienia zadania), a w polu **Plik główny** wskazujemy plik ze skryptem.



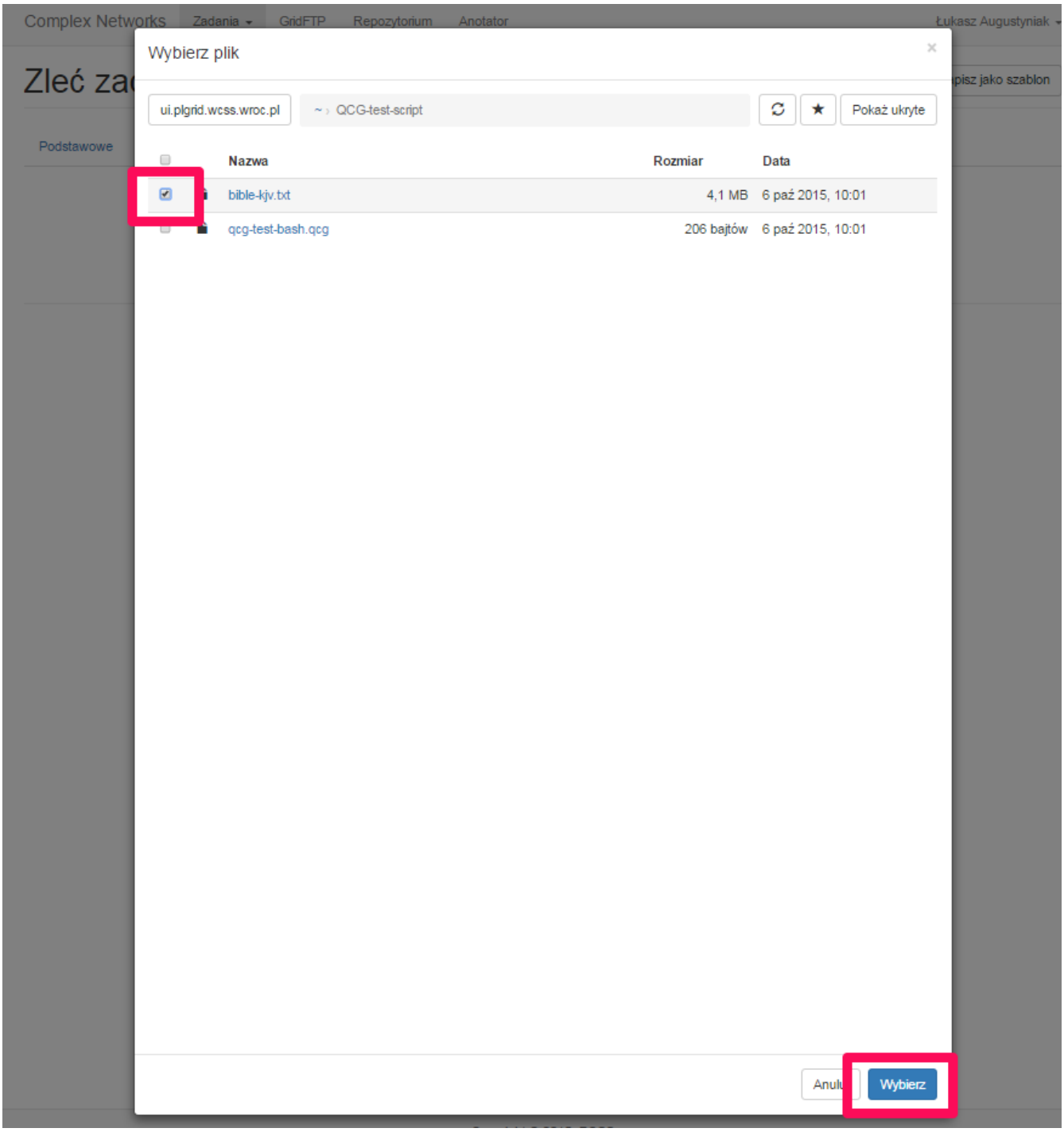
Zaznaczamy wybór przy nazwie pliku i potwierdzamy przyciskiem **Wybierz.**



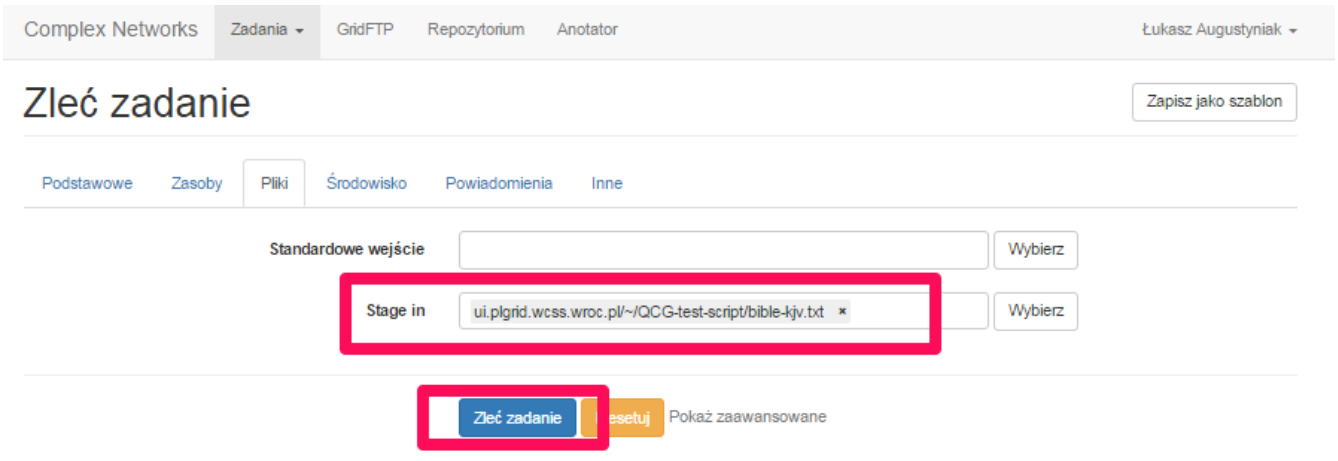
Następnie dodamy do zadania plik tekstowych, dla którego policzymy liczbę słów w nim występujących - zgodnie ze skryptem BASH. W tym celu klikamy na zakładkę **Plik** oraz wybieramy atrybut **Stage in.**



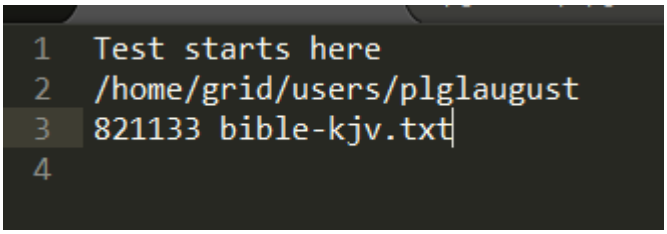
Wskazujemy plik podobnie jak przy skrypcie bash.



W tym momencie przykładowe zadanie jest ustawione i można je zlecić przyciskiem **Zleć zadanie.**



W konsekwencji otrzymujemy wynik jak poniżej (podgląd pliku wyjściowego) - plik można znaleźć wybierając **Zadania -> Moje zadania** i wchodząc w **s zczegóły** zleconego zadania oraz wybierając **Przeglądaj katalog roboczy**.



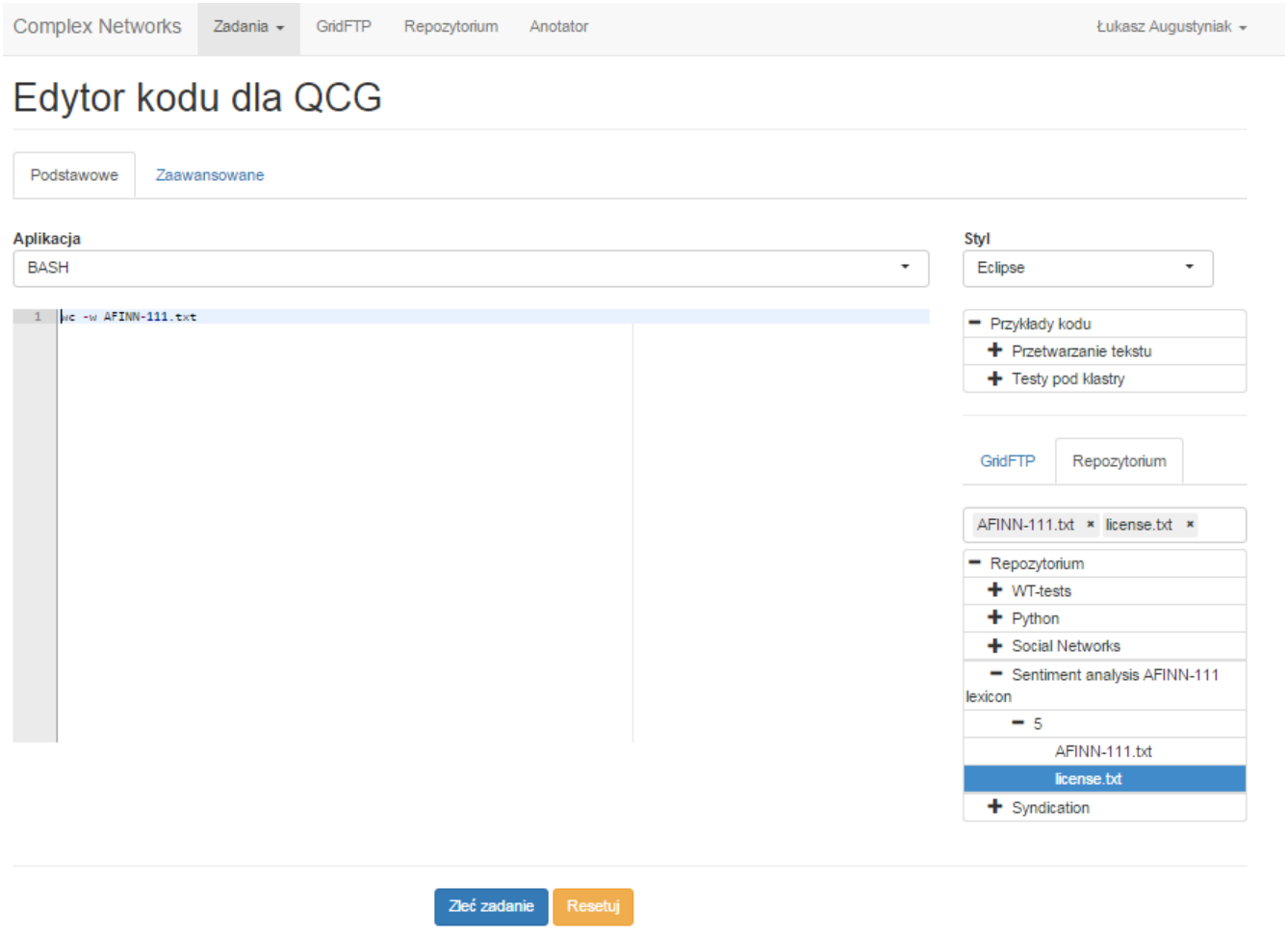
Zlecanie zadań z wykorzystaniem Edytora QCG

Widok składa się co do zasady z 3 części:

1. Edytor kodu BASH/Python (lewa część interfejsu)
2. Sekcja wyboru przykładowych kodów (prawa górna część)
3. Sekcja wyboru zbioru danych (prawa dolna część)

Część pól z przedstawionego powyżej formularza pozostaje także i w tym widoku. Umożliwiają one dostosowanie zadania do naszych potrzeb. Bardzo ważnym jest wybór Aplikacji, tj. BASH albo Python, gdyż w ten sposób określamy jakiego rodzaju kod może zostać użyty do uruchomienia zadania.

W przykładzie poniżej widać wywołanie kodu BASH **wc -w AFINN-111.txt.** Należy zwrócić uwagę na wskazanie pliku **AFINN-111.txt,** plik ten został wybrany z repozytorium (po zleceniu zadania zostanie pobrany z repozytorium i przeniesiony na maszyny obliczeniowe - nie musimy nic dodatkowo kopiować!) Jeśli chcemy przenieść więcej plików należy je odnaleźć w repozytorium, a następnie pojedynczym kliknięciem myszy dodać do zadania. W celu dodania pliku w kodzie należy (po dodaniu go do zadania) ustawić kursor w odpowiednim miejscu kodu oraz kliknąć dwukrotnie na nazwę pliku (w poniższym przypadku można wybrać `AFINN-111.txt` albo `license.txt` w sekcji repozytorium). Usuwanie plików z zadania - kliknięcie krzyżyka po prawej stronie nazwy pliku.



W celu wykonania zadania należy opakować kod Python w następującą strukturę:

python - <<END

kod python

END

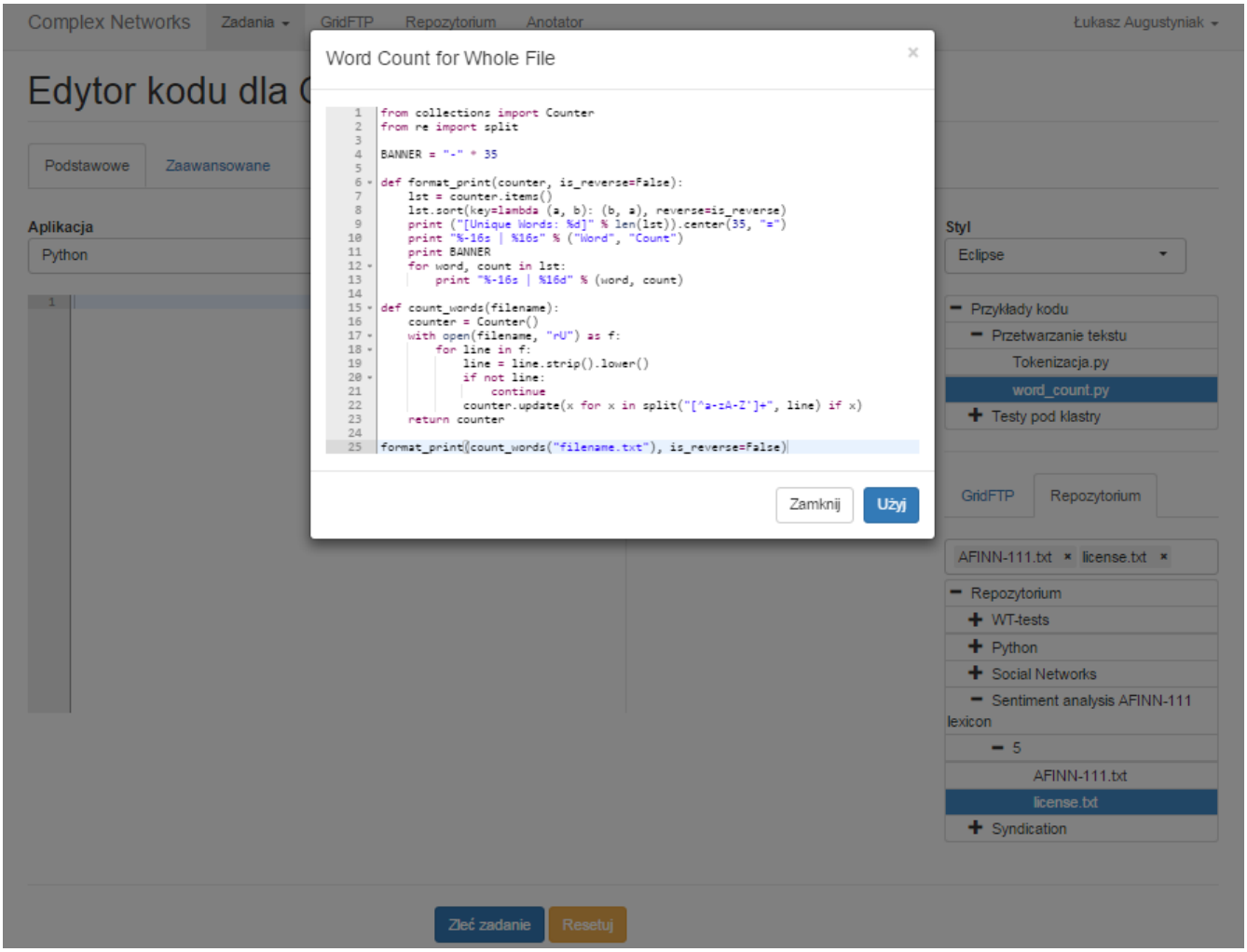
Patrz przykład z przetwarzaniem danych sieciowych poniżej

Zlecanie zadania z wykorzystaniem przykładowych kodów - Python

Chcąc zlecić zadanie w języku Python należy zmienić typ Aplikacji z **BASH** na **Python.**

Następnie wyszukać interesujący nas fragment kodu, w przykładzie kod obliczający liczbę wystąpień poszczególnych słów w tekście **- word_count.py**

W celu podglądu zawartości kodu należy kliknąć na jego nazwę, w ten sposób dostaniemy dodatkowe okienko z podglądem.

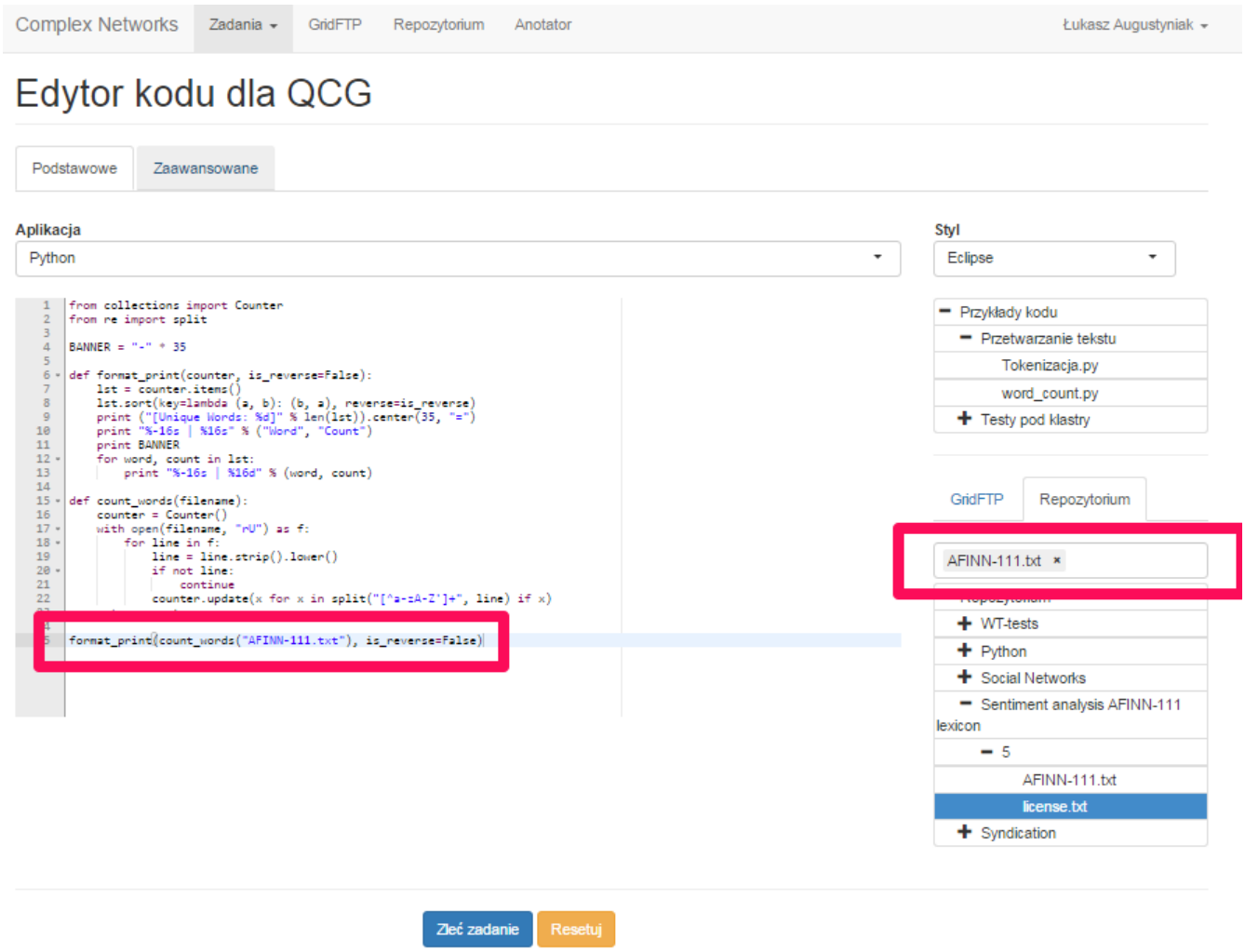


Wybieramy przycisk **Użyj,** a kod zostanie skopiowany do Edytora.

Następnie musimy do zadania dodać z repozytorium plik, dla którego chcemy wykonać analizę oraz zmodyfikować kod Python w taki sposób, żeby wskazać tenże plik.

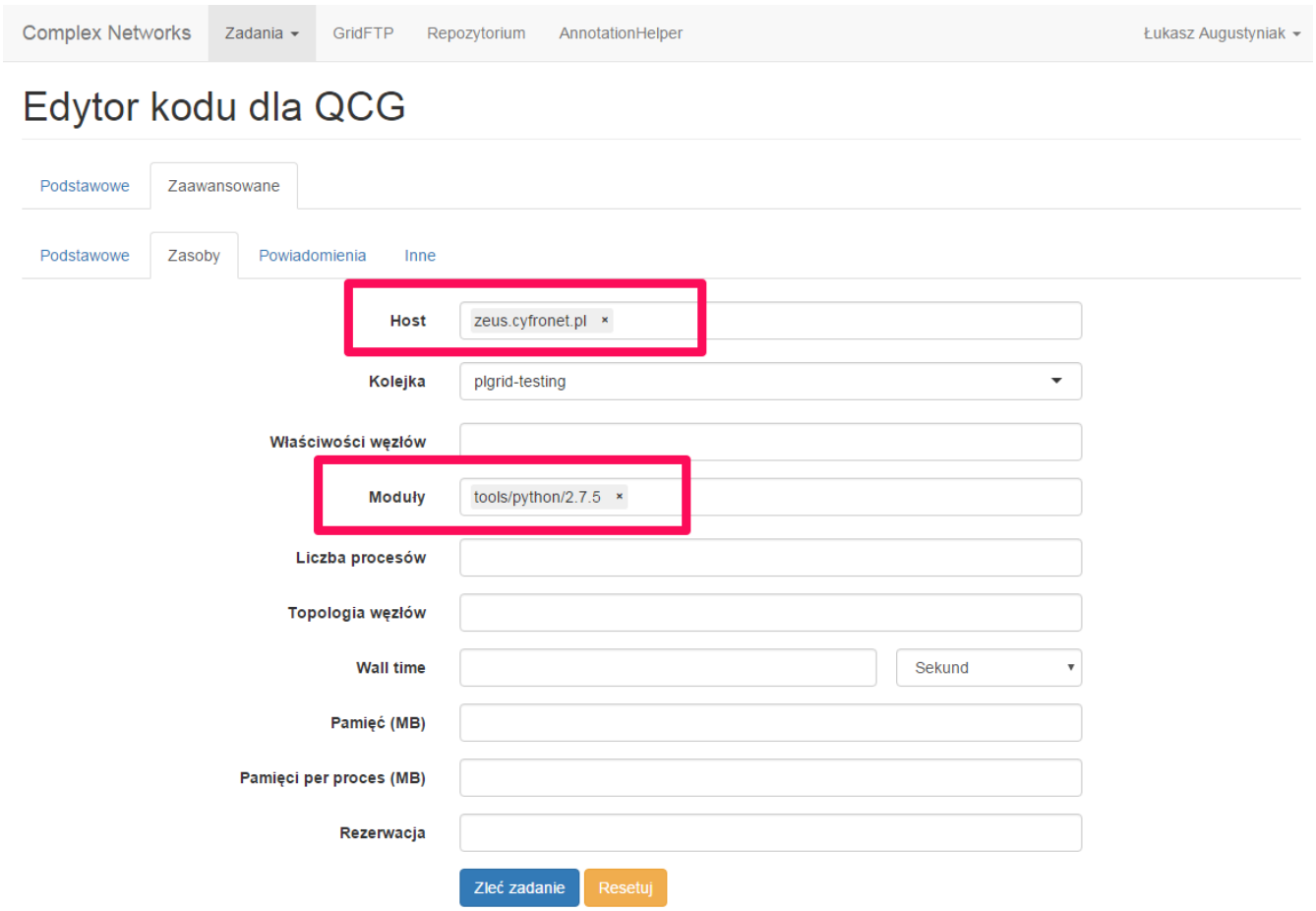
W tym celu wyszukujemy w repozytorium interesujący nas plik, dodajemy go do zadania (tak samo jak wybieraliśmy podgląd kodu przykładowego), a następnie wskazujemy miejsce w kodzie gdzie chcemy dodać plik i klikamy dwukrotnie na jego nazwę (prawe zaznaczenie na poniższym zrzucie ekranu).

W efekcie widzimy zmodyfikowany kod w Edytorze.



Warto zwrócić uwagę na zakładkę **Zaawansowane,** gdzie to możemy określić parametry wykonania zadania zgodnie z Format opisu zadań. W przypadku zlecania zadań Python jest to bardzo ważne, ponieważ musimy wybrać odpowiedni moduł (z preinstalowanymi bibliotekami Python dla gridu Complex Networks).

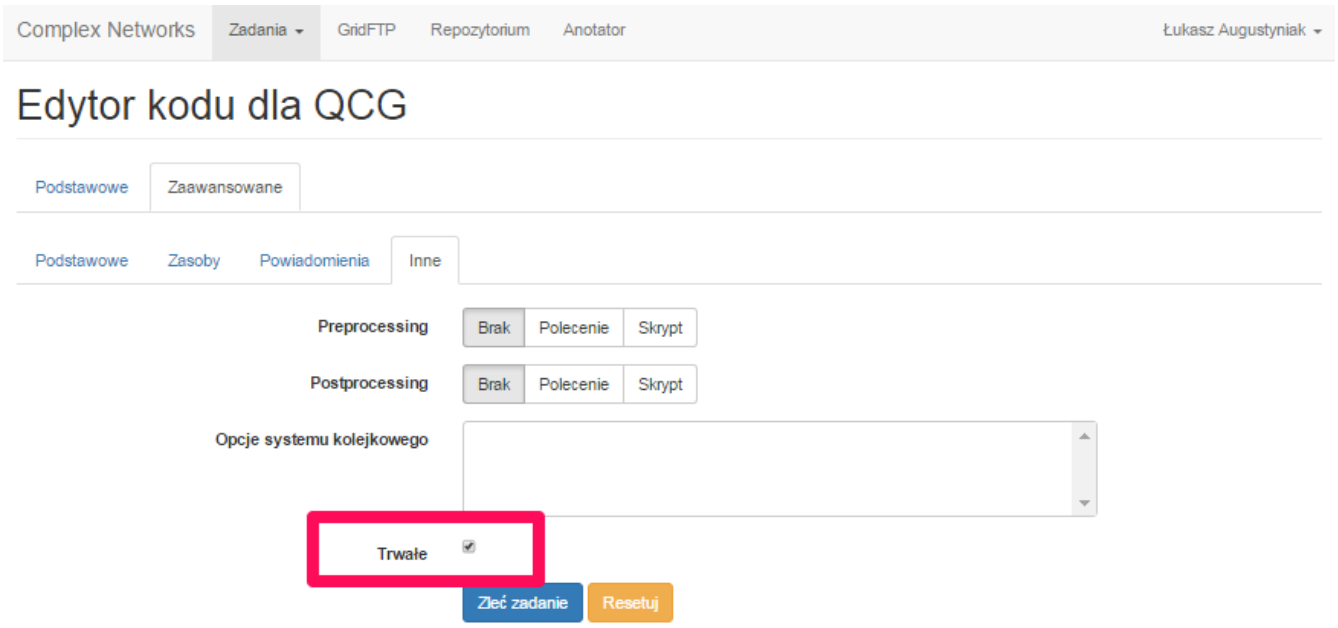
W przedstawionym poniżej przypadku zlecamy zadanie obliczeniowe dla klastra **Zeus,** a to pociąga wybór modułu Python odpowiedniego właśnie dla tego klastra. W przypadku wyboru innego/dodatkowego klastra należy uwzględnić to w parametrze **Moduły.**



Odpowiednie biblioteki Python są już pre-instalowane w następujących modułach:

Python-Task - plgrid/tools/python/2.7.9
Python-Zeus - tools/python/2.7.5
Python-Bem - plgrid/tools/python/2.7.8
Python-Inula - plgrid/tools/python/2.7.3

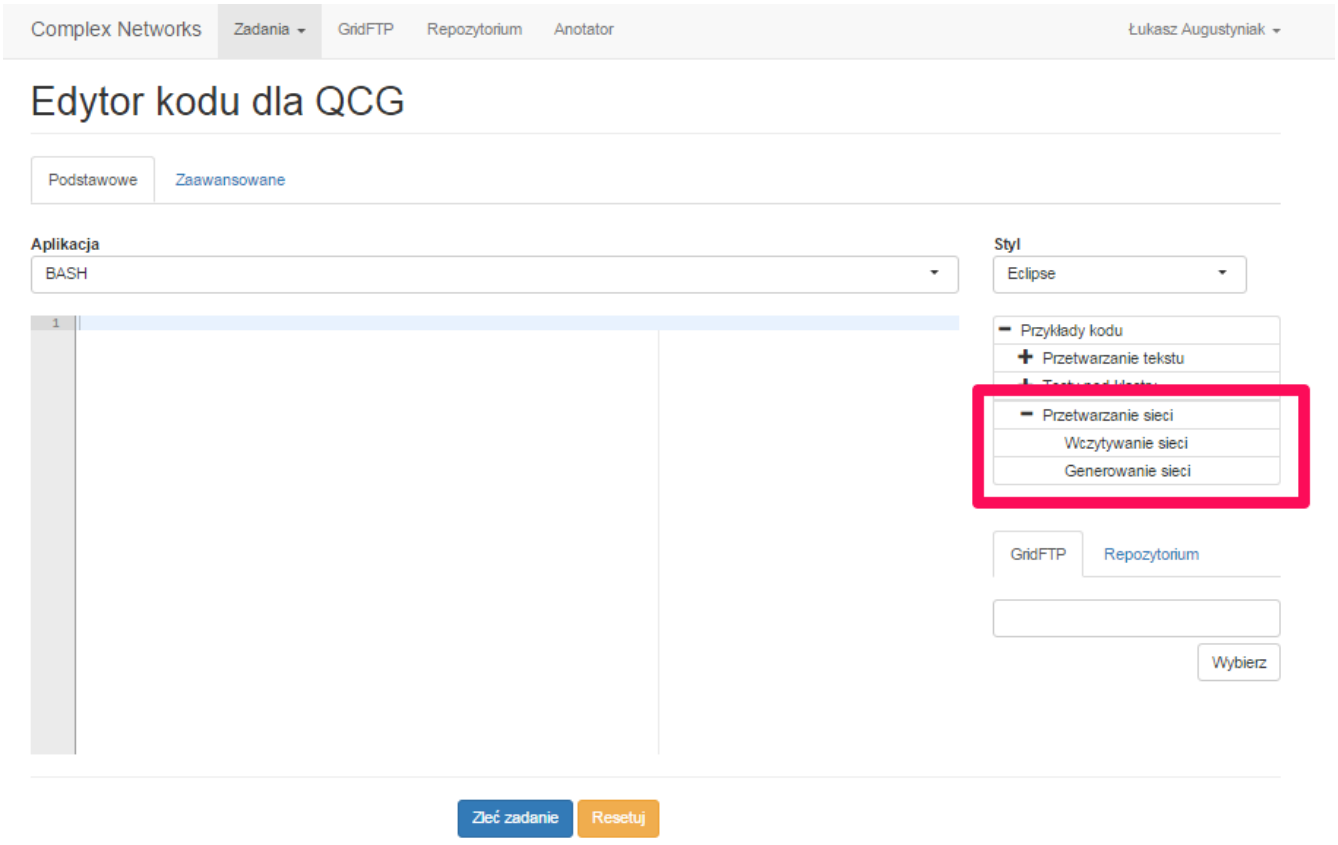
UWAGA!! Proszę nie usuwać zaznaczenia **TRWAŁE** z formularza specyfikacji parametrów zadania. W przeciwnym razie zadania zostaną wykonane na klastrze, ale nie zostaną zachowane i skopiowane z powrotem na maszyny lokalne, przez co nie będzie można zobaczyć wyników analiz!

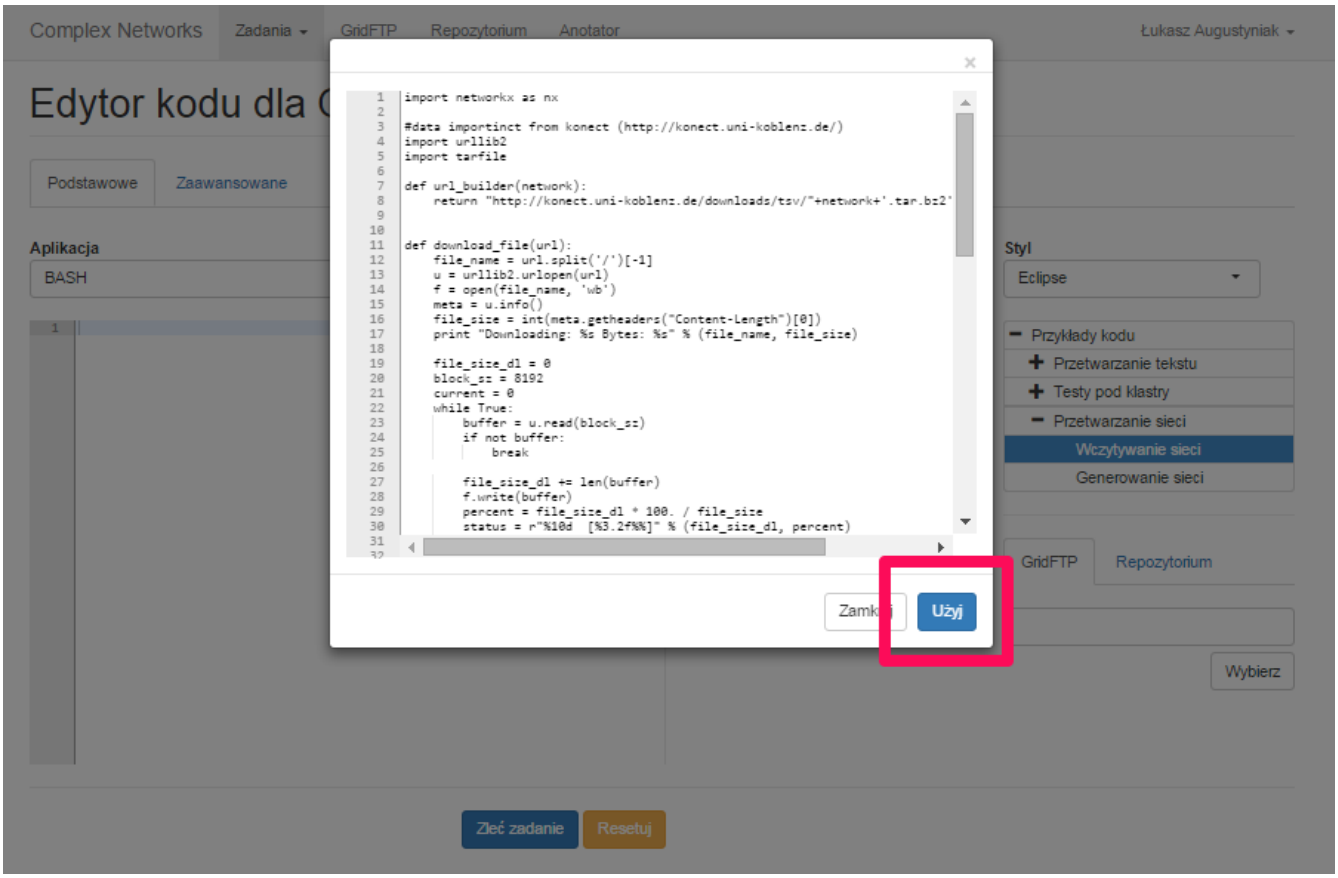


Zaawansowane użycie

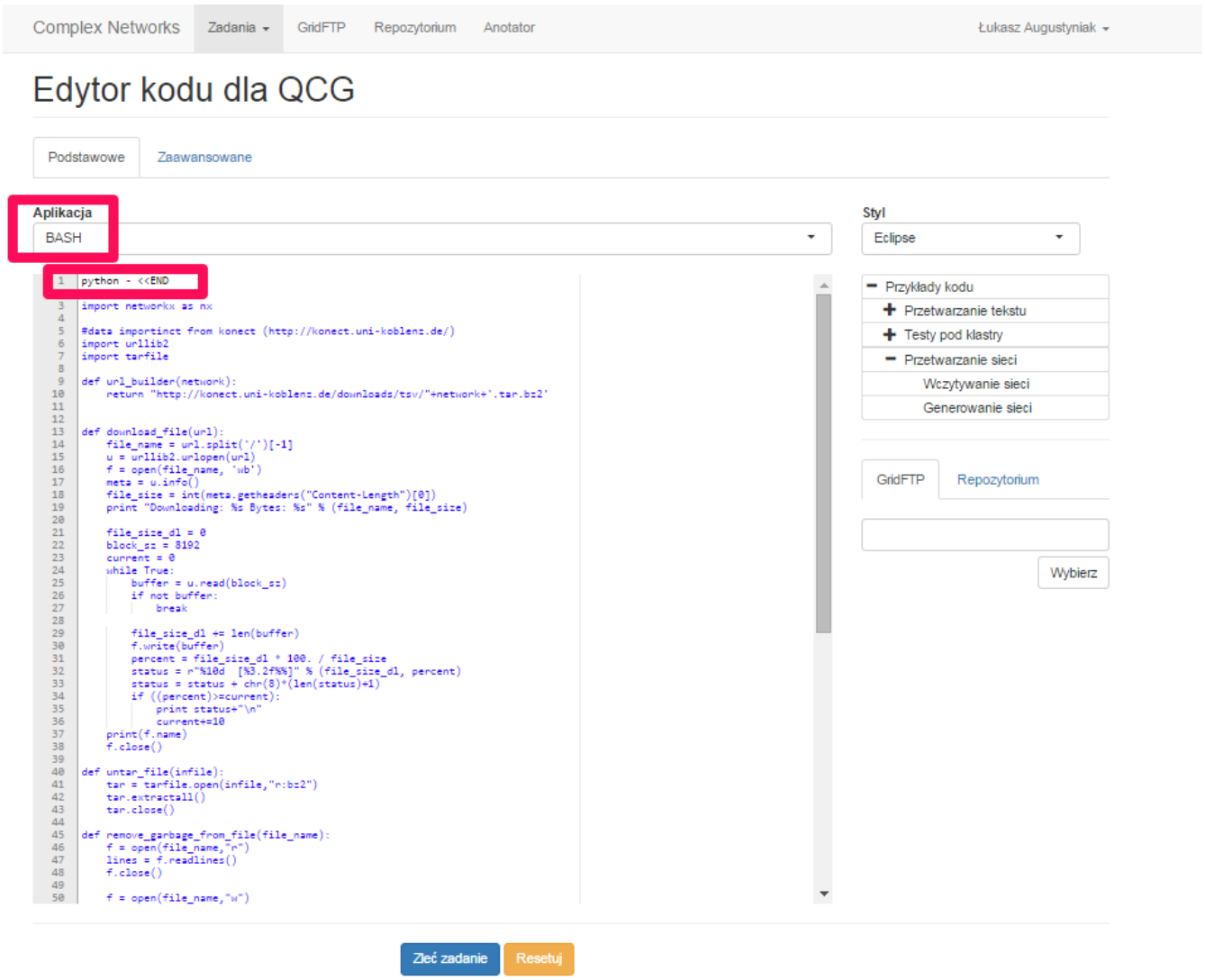
Zlecanie zadania Python - wczytywanie danych sieciowych

W **Przykładach kodu** wybieramy **Przetwarzanie sieci**, a następnie **wczytywanie sieci.**

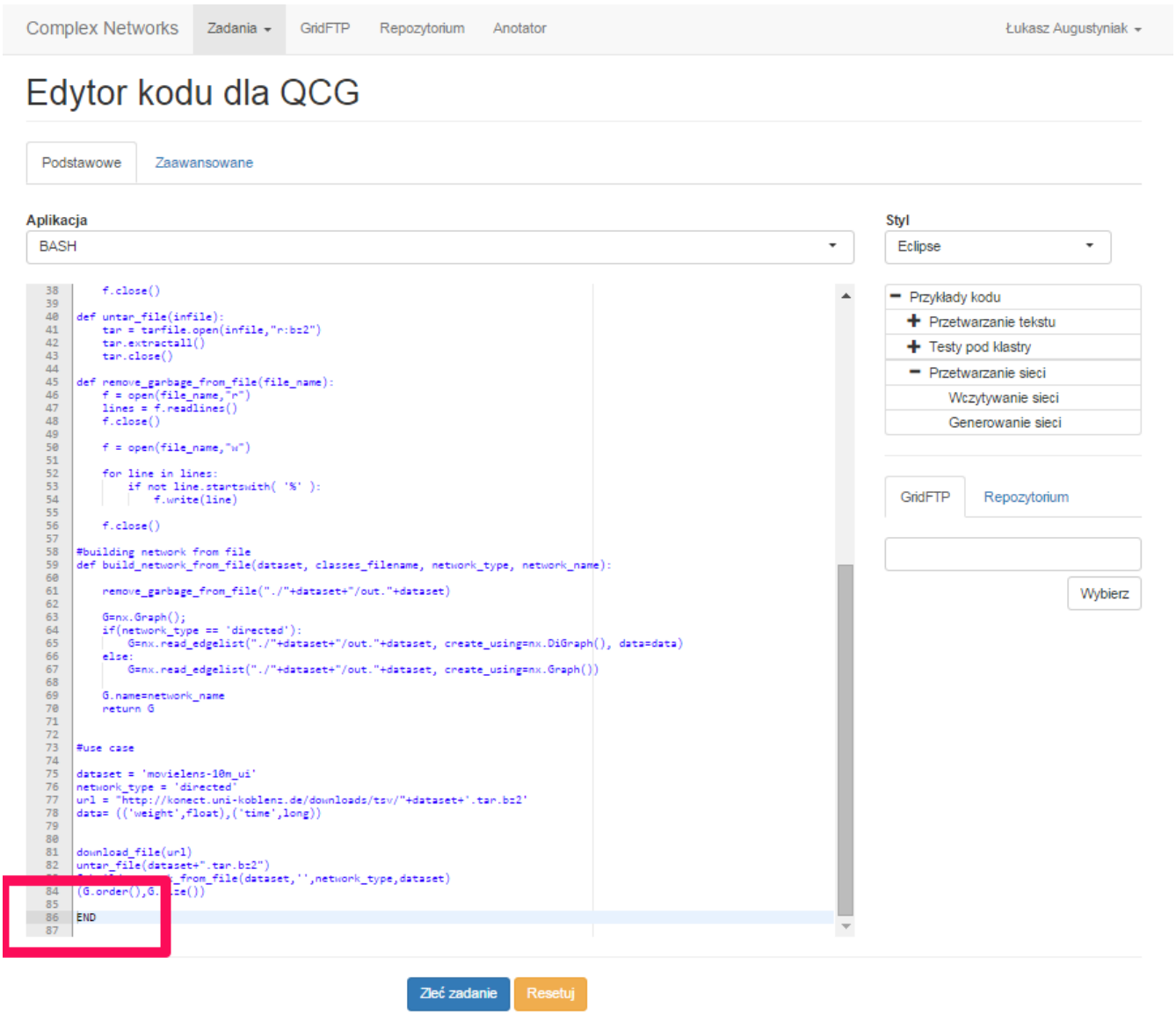




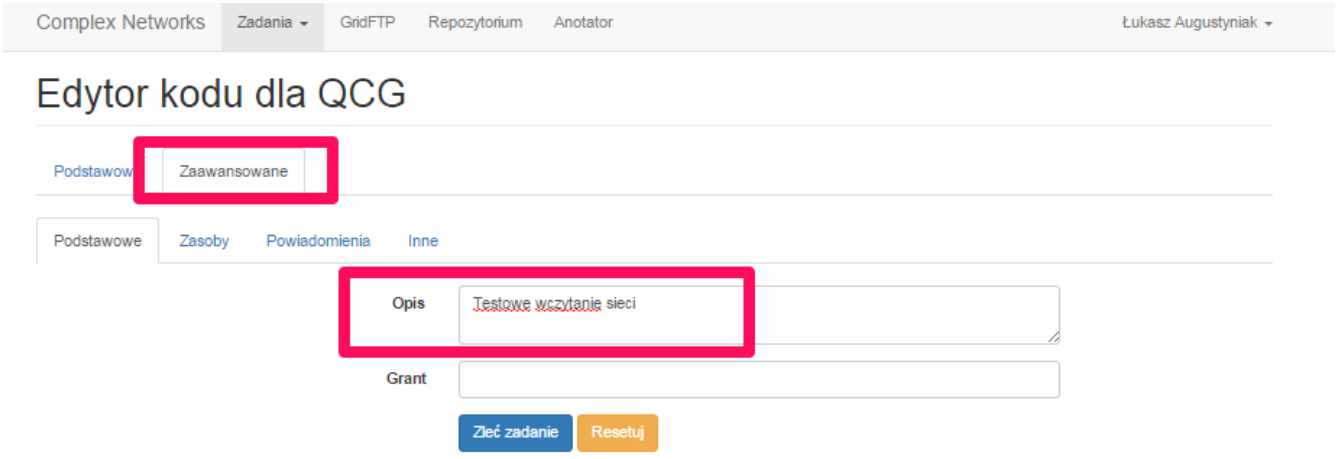
Dla klastrów innych niż Bem proszę dopisać kod **python - <<END** na początku skryptu, tak jak na poniższym przykładzie.



oraz **END** na końcu kodu.



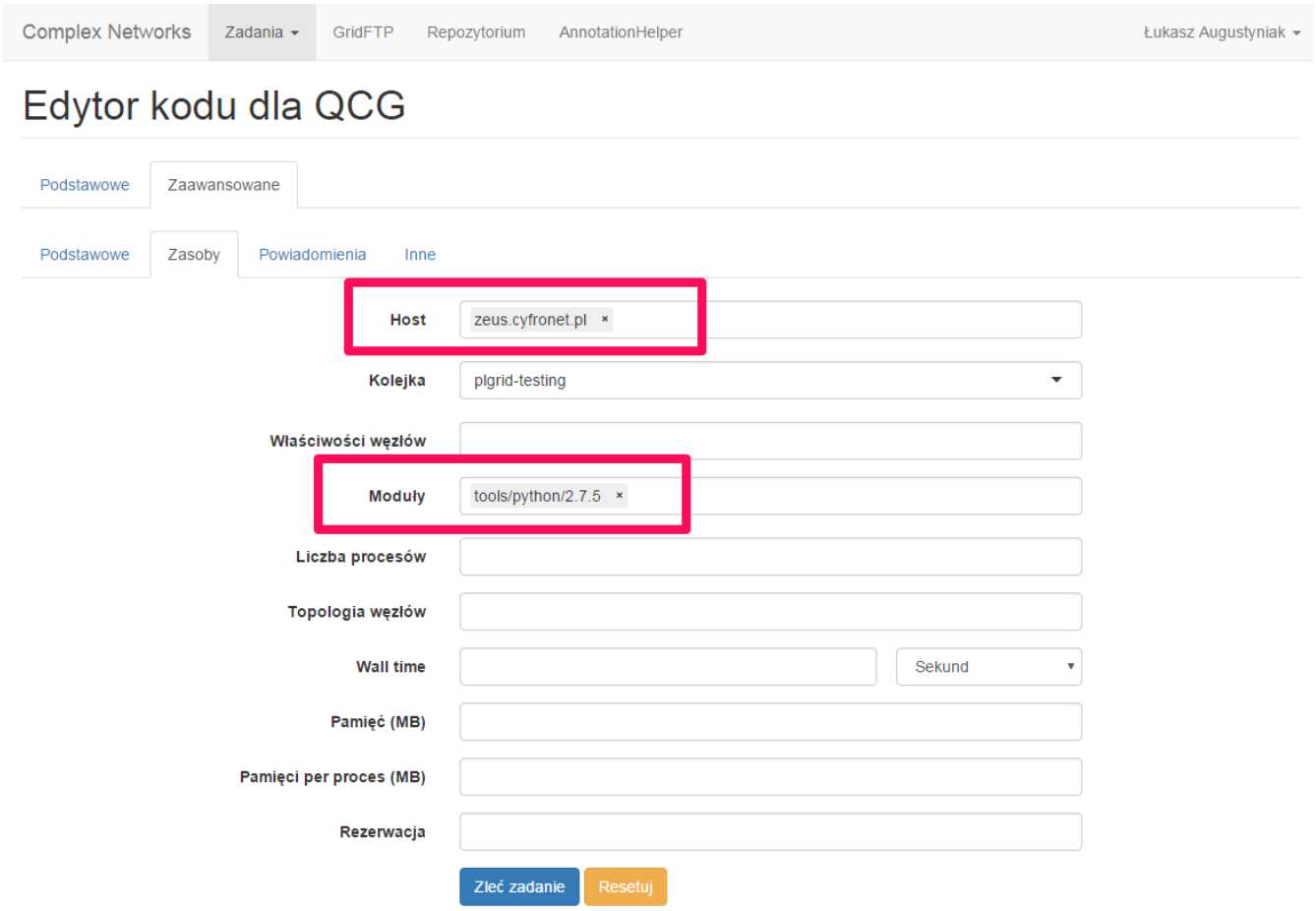
Następnie określamy opis zlecanego zadania w zakładce **Zaawansowane -> Opis.**



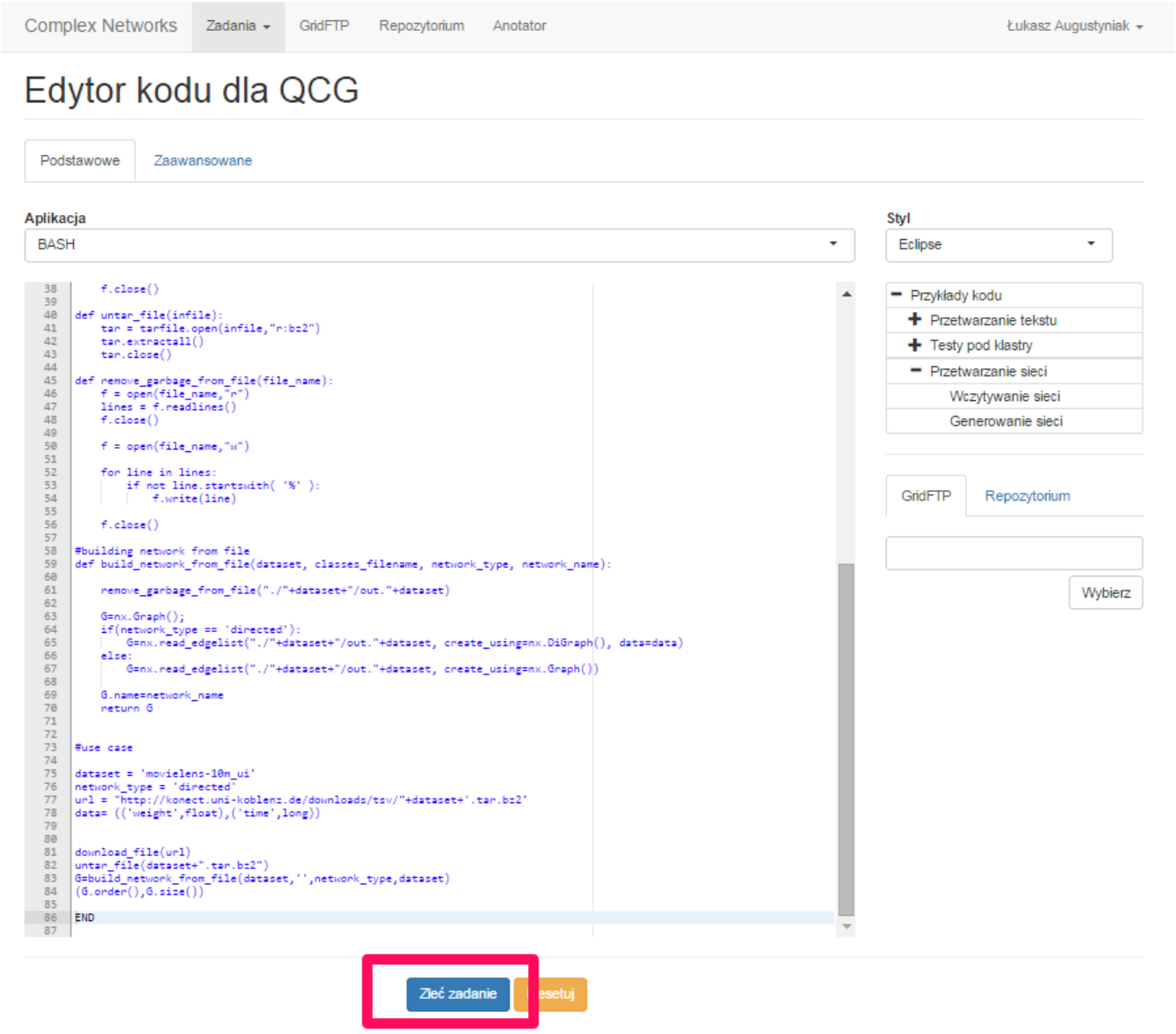
Wybieramy **Host (klaster)** oraz odpowiedni moduł Python dla klastra.

Odpowiednie biblioteki Python są już pre-instalowane w następujących modułach:

Python-Task - plgrid/tools/python/2.7.9
Python-Zeus - tools/python/2.7.5
Python-Bem - plgrid/tools/python/2.7.8
Python-Inula - plgrid/tools/python/2.7.3



Wszystko gotowe i można zlecić zadanie.




```
    f.close()

def untar_file(infile):
    tar = tarfile.open(infile,"r:bz2")
    tar.extractall()
    tar.close()

def remove_garbage_from_file(file_name):
    f = open(file_name,"r")
    lines = f.readlines()
    f.close()

    f = open(file_name,"w")

    for line in lines:
        if not line.startswith( '%' ):
            f.write(line)

    f.close()

#building network from file
def build_network_from_file(dataset, classes_filename, network_type, network_name):

    remove_garbage_from_file("./"+dataset+"/out."+dataset)

    G=nx.Graph();
    if(network_type == 'directed'):
        G=nx.read_edgelist("./"+dataset+"/out."+dataset, create_using=nx.DiGraph(), data=data)
    else:
        G=nx.read_edgelist("./"+dataset+"/out."+dataset, create_using=nx.Graph())

    G.name=network_name
    return G

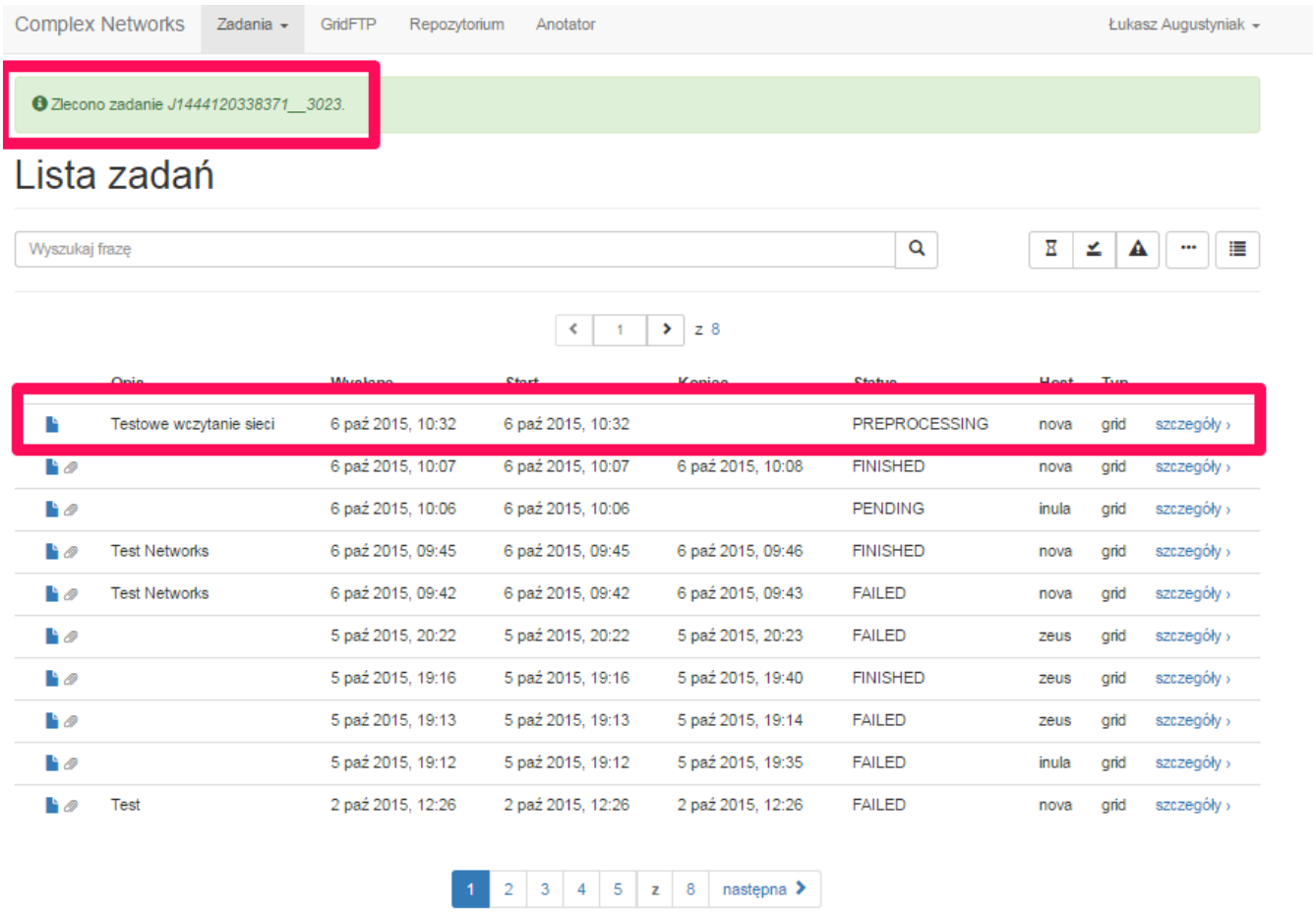

#use case

dataset = 'movielens-10m_ui'
network_type = 'directed'
url = "http://konect.uni-koblenz.de/downloads/tsv/"+dataset+".tar.bz2"
data= (('weight',float),('time',long))


download_file(url)
untar_file(dataset+".tar.bz2")
G=build_network_from_file(dataset,'',network_type,dataset)
(G.order(),G.size())

END
```

Zostajemy poinformowani o nazwie zlecanego zadania oraz przeniesieni do podglądu naszych wszystkich zlecanych zadań.

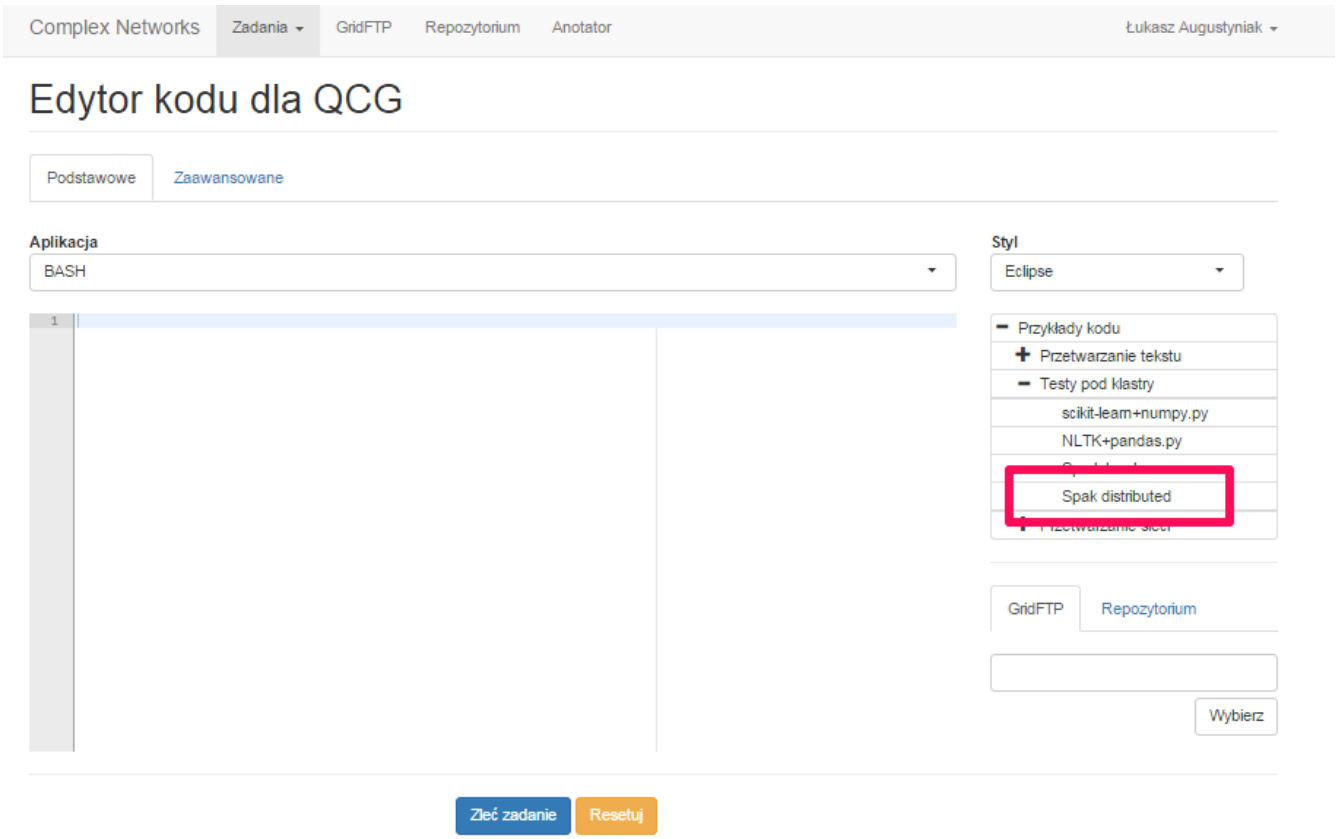


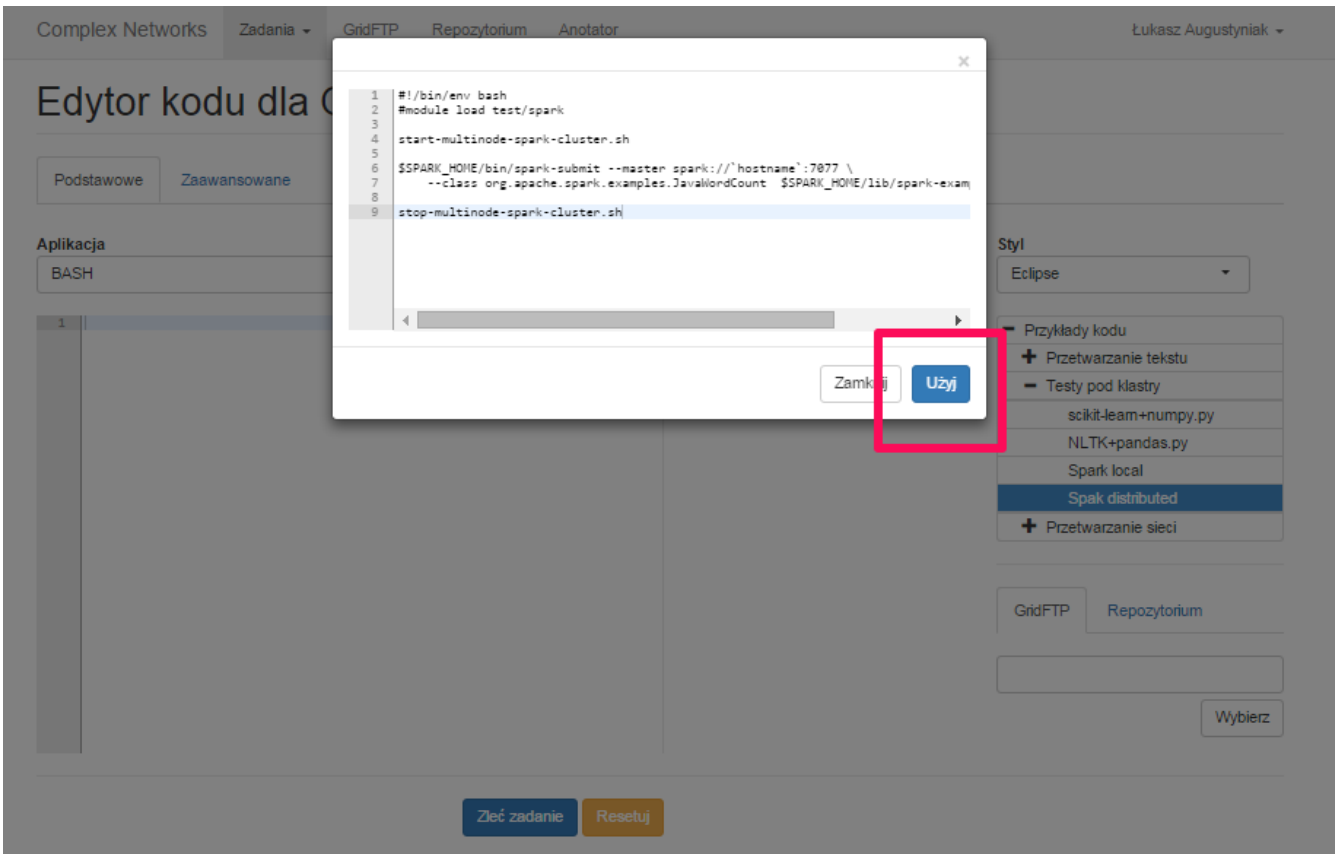
Zlecanie zadania SPARK

Wchodzimy do zakładki **Zadania/Edytor QCG**

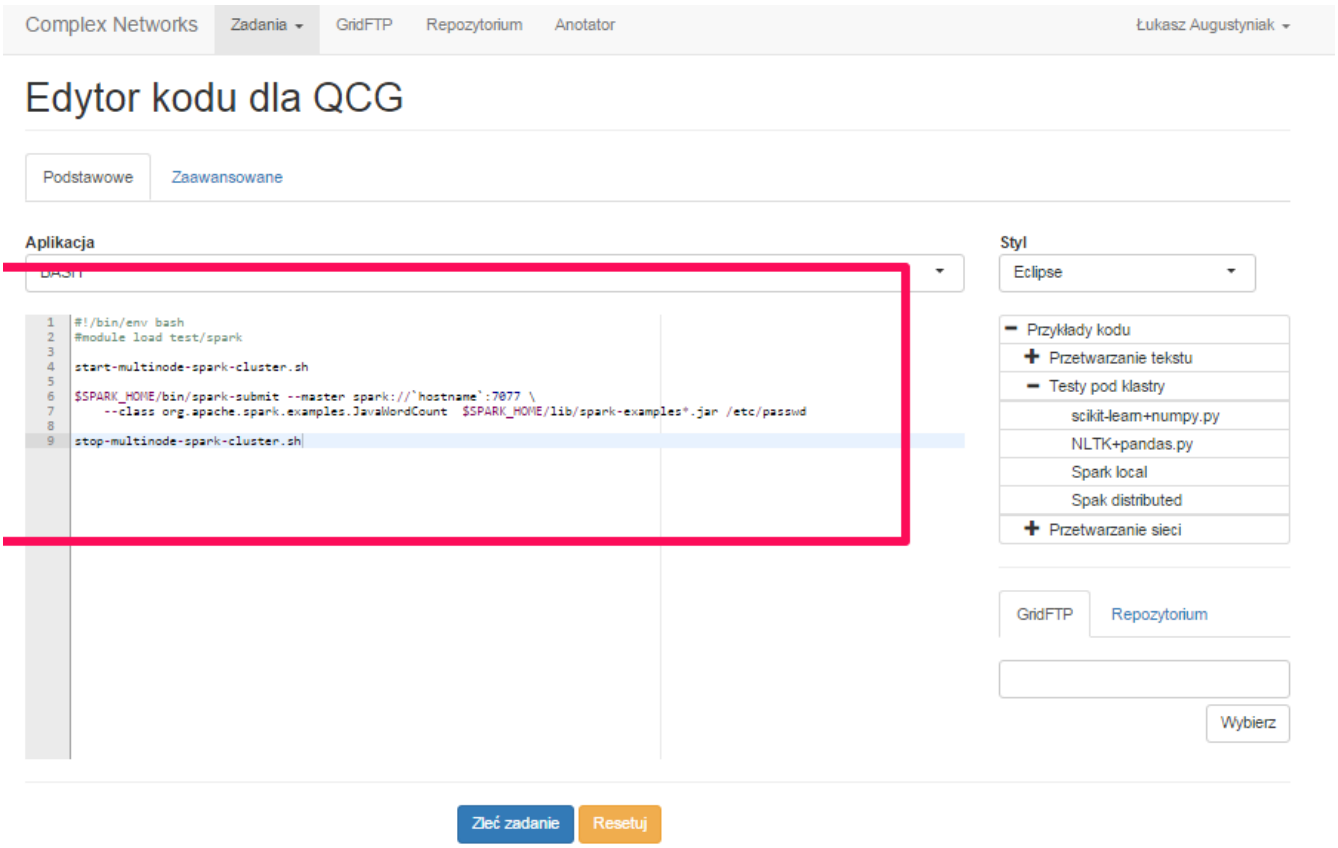
W polu Aplikacja wybieramy **bash**

Następnie wybieramy Przykład kodu **Spark distributed**

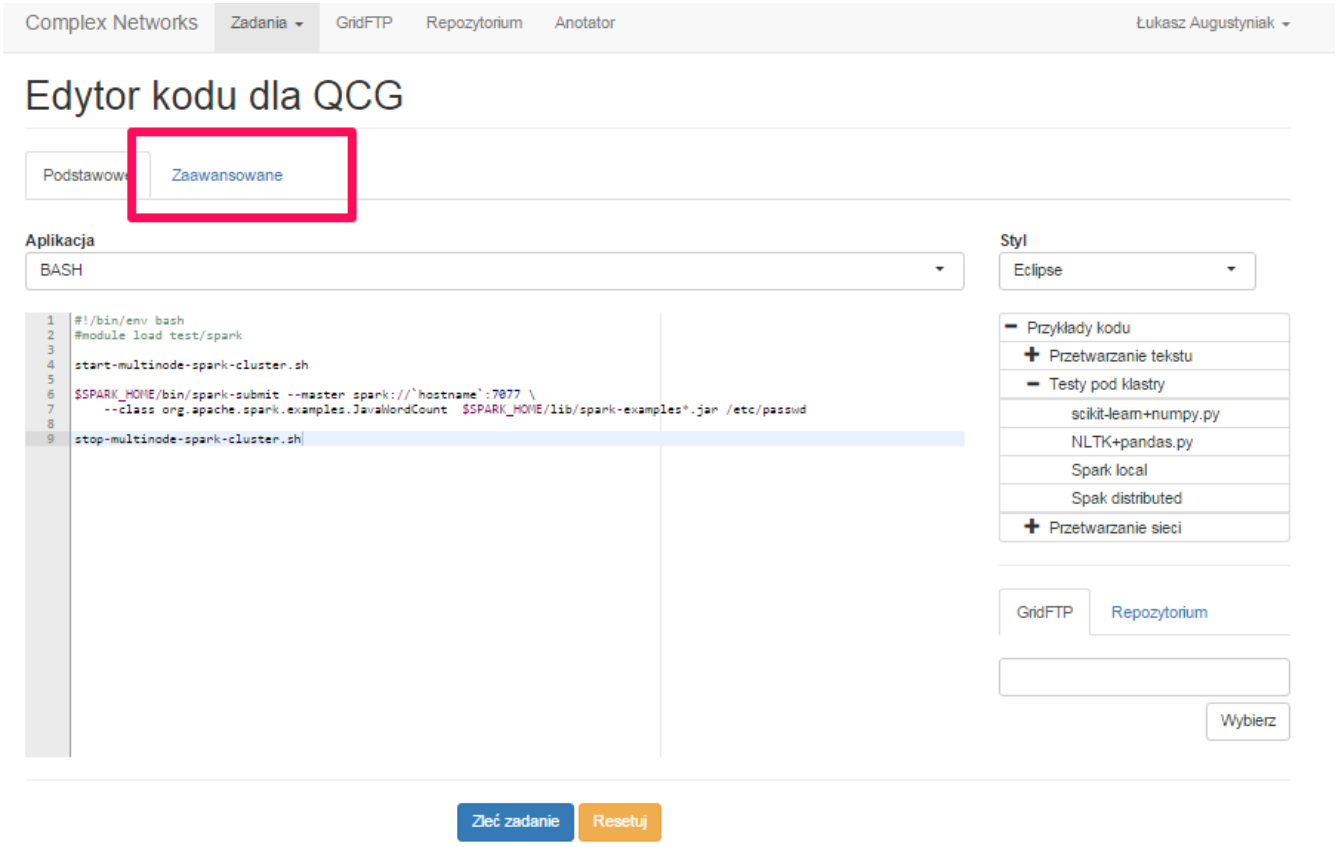




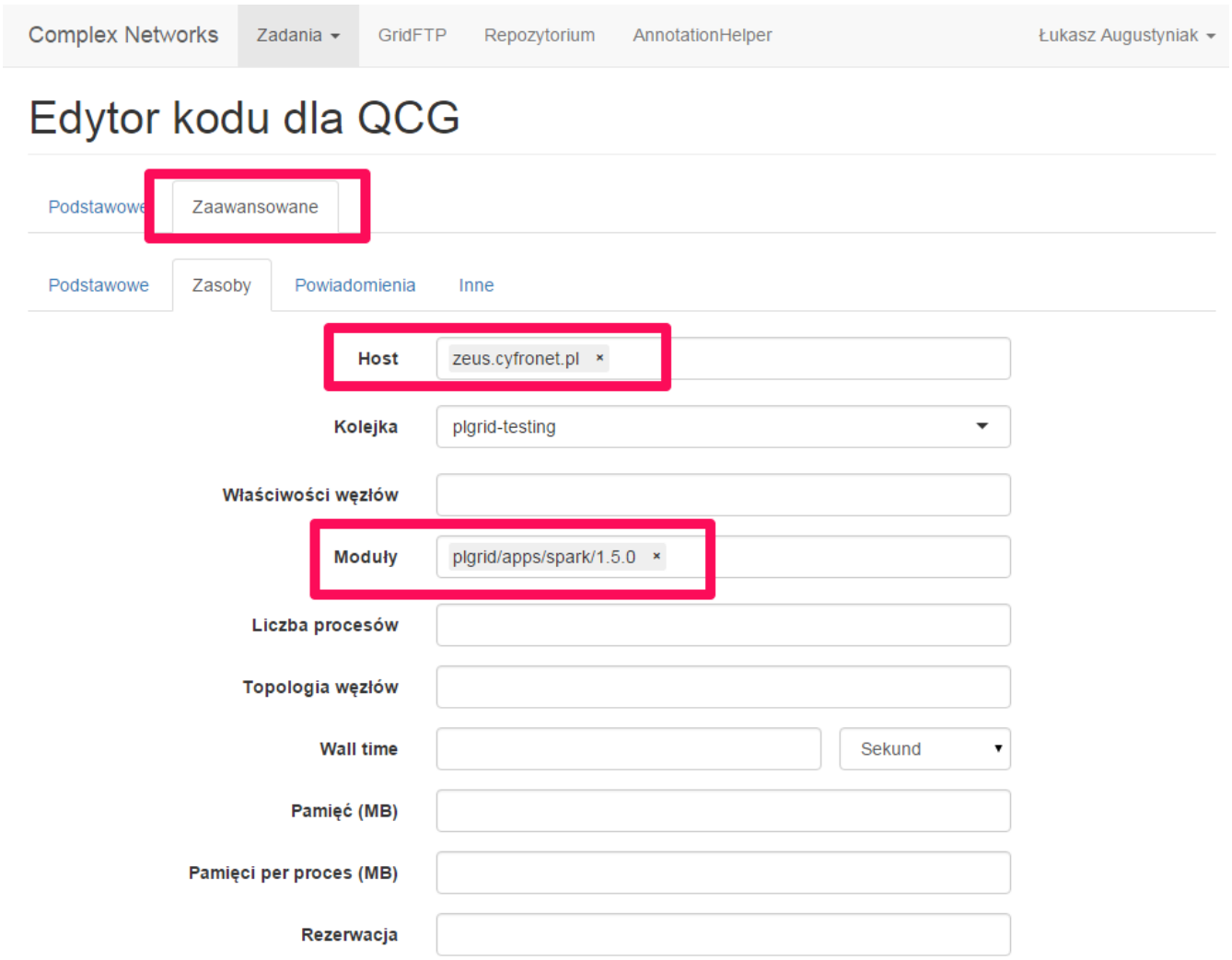
Na poniższym zrzucie widać przykładowy kod wywołujący zadanie z użyciem framework Spark - przykład zliczania słów w sposób rozproszony



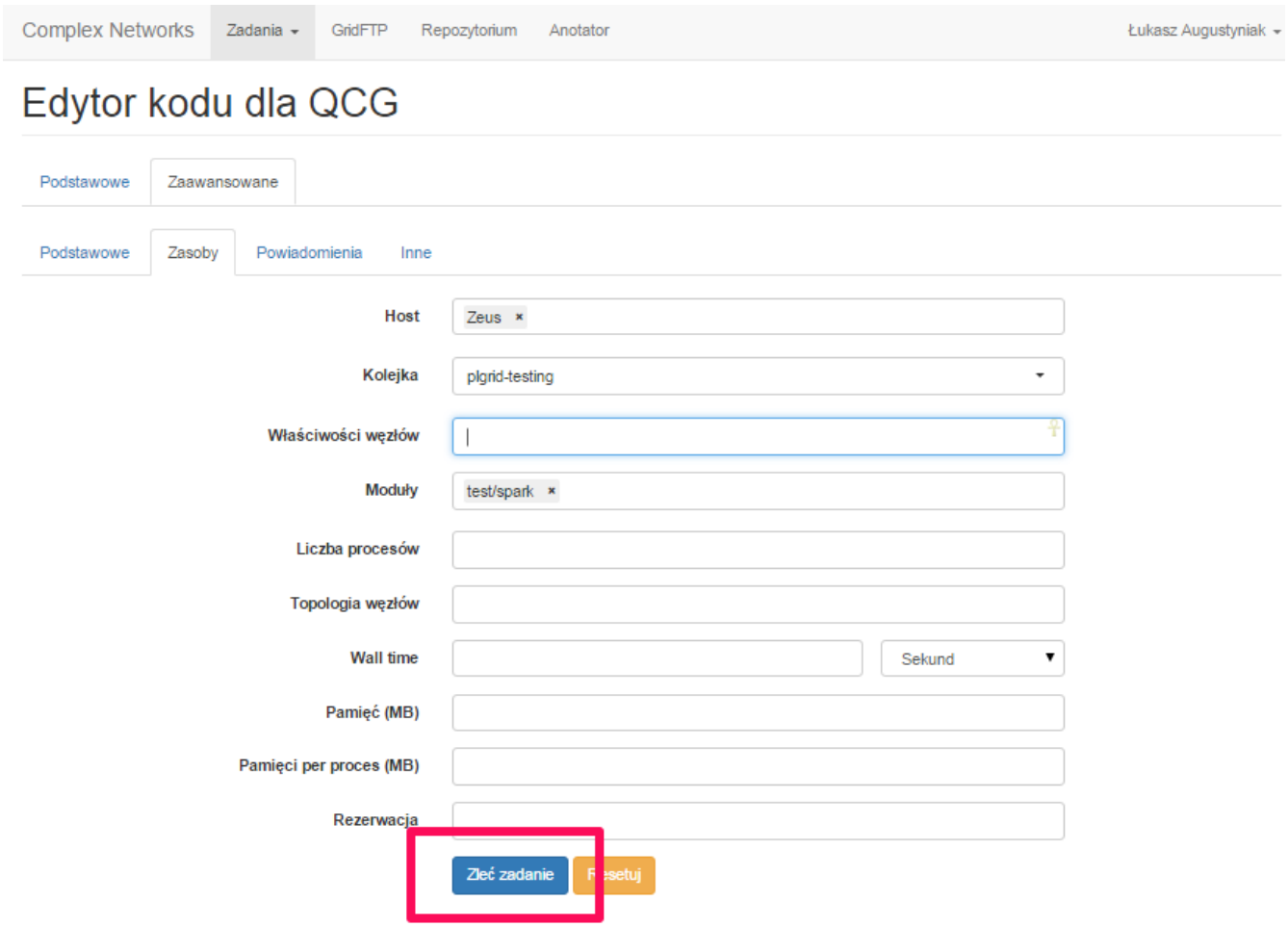
Następnie musimy wybrać odpowiedni klaster obliczeniowy (niestety nie wszystkie wspierają Spark). Przechodzimy do zakładki **Zaawansowane**



Dalej wybieramy klaster **Zeus** oraz moduł **Spark**



W tym momencie mamy ustawione wszystkie niezbędne parametry do wywołania zadania. Możemy zlecić zadania klikając na przycisk **Zleć zadanie**



Status zadania możemy podejrzeć w zakładce **Zadania/Moje zadania**. Widzimy tutaj nazwę, numer, czas zlecenie oraz inne informacje związane z naszym zadaniem. Dodatkowo wybierając **szczegóły** możemy zobaczyć szczegółowe informacje dotyczące naszego zadania oraz wynik jego wykonania.

Complex Networks | Zadania | GridFTP | Repozytorium | Anotator | Łukasz Augustyniak

Zlecono zadanie *J1444069350161__9718.*

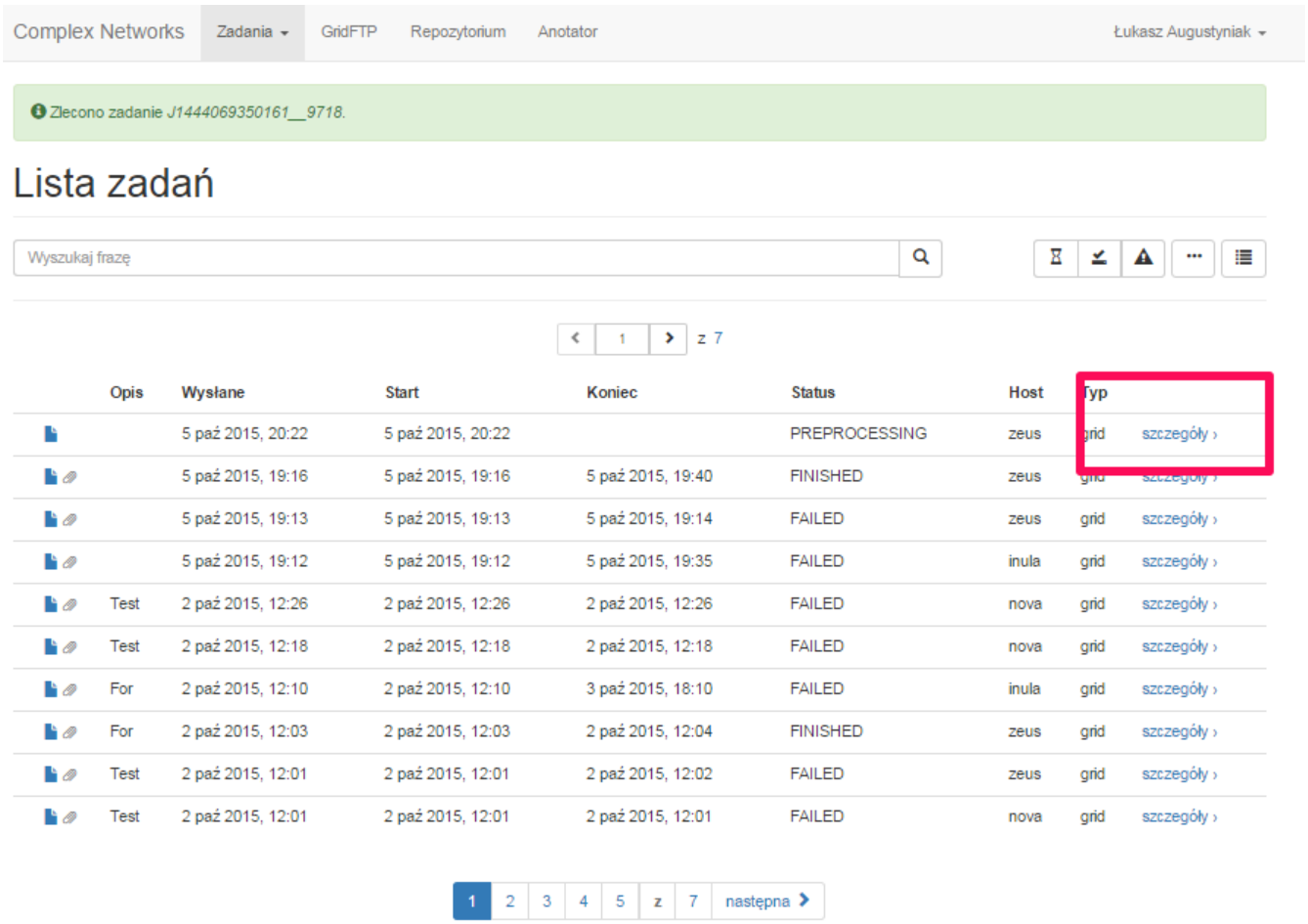
Lista zadań

Wyszukaj frazę

1 z 7

	Opis	Wysłane	Start	Koniec	Status	Host	Typ	
		5 paź 2015, 20:22	5 paź 2015, 20:22		PREPROCESSING	zeus	grid	szczegóły ›
		5 paź 2015, 19:16	5 paź 2015, 19:16	5 paź 2015, 19:40	FINISHED	zeus	grid	szczegóły ›
		5 paź 2015, 19:13	5 paź 2015, 19:13	5 paź 2015, 19:14	FAILED	zeus	grid	szczegóły ›
		5 paź 2015, 19:12	5 paź 2015, 19:12	5 paź 2015, 19:35	FAILED	inula	grid	szczegóły ›
	Test	2 paź 2015, 12:26	2 paź 2015, 12:26	2 paź 2015, 12:26	FAILED	nova	grid	szczegóły ›
	Test	2 paź 2015, 12:18	2 paź 2015, 12:18	2 paź 2015, 12:18	FAILED	nova	grid	szczegóły ›
	For	2 paź 2015, 12:10	2 paź 2015, 12:10	3 paź 2015, 18:10	FAILED	inula	grid	szczegóły ›
	For	2 paź 2015, 12:03	2 paź 2015, 12:03	2 paź 2015, 12:04	FINISHED	zeus	grid	szczegóły ›
	Test	2 paź 2015, 12:01	2 paź 2015, 12:01	2 paź 2015, 12:02	FAILED	zeus	grid	szczegóły ›
	Test	2 paź 2015, 12:01	2 paź 2015, 12:01	2 paź 2015, 12:01	FAILED	nova	grid	szczegóły ›

1 2 3 4 5 z 7 następna ›



Gdzie szukać dalszych informacji?

Więcej informacji można znaleźć na oficjalnej stronie usługi https://cn.plgrid.pl. Pytania proszę zadawać poprzez: https://helpdesk.plgrid.pl

Dodatkowo polecamy stronę Helpdesk PLGrid