

AnnotationHelper

Ze względu na jego objętość podręcznik użytkownika jest również dostępny jako plik pdf: [Podręcznik Użytkownika platformy Annotation Helper v1.2.pdf](#)

Dodatkowo: [Przypadek użycia "comedy club"](#), Zbiór danych: [dowcipy.csv](#)

Niezbędne informacje wstępne

W ramach gridu dziedzinowego Complex Networks, zrealizowanego w ramach projektu PIGridNG, wspierane są trzy usługi:

1. Usługa syndykacji danych - „Syndykacja” – usługa wspierająca użytkowników w zbieraniu danych z Web 2.0
2. Usługa analiz danych tekstowych i sieciowych – „Analizy”
3. Usługa udostępniania zbiorów danych – „Repozytorium”

Platforma AnnotationHelper jest traktowana jako część usługi syndykacji danych, która zakłada, że użytkownicy (ludzie) są częścią Internetu. Na tej zasadzie ich wiedza jest również zasobem Internetu. Zbieranie adnotacji od użytkowników Internetu również jest zbieraniem danych z Internetu.

Celem AnnotationHelper jest wsparcie naukowców z różnych dziedzin w dokonywaniu adnotacji (etykietowaniu) multimedialnych zbiorów danych (dowolna kombinacja tekstu, obrazków, muzyki i video).

Przykładowe zastosowania to: etykietowanie zdjęć pobranych z mediów społecznościowych pod kątem ich treści, etykietowanie tekstów pod kątem ich wydźwięku, emocji w nich zawartych lub emocji wzbudzających, rozpoznawanie dźwięków w pliku muzycznym lub video i inne... Etykietowanie danych jest rzeczą niezbędną m.in. w trakcie tworzenia zbioru uczącego w nadzorowanym uczeniu maszynowym (ang. *supervised machine learning*).

Platforma ta jest ściśle związana z „Repozytorium”, z którą wymienia dane. Z niej pobiera zbiór do przygotowania jego adnotacji, a po zakończeniu zwraca rozszerzony o adnotację zbiór z powrotem do repozytorium.

Wgranie zbioru danych do repozytorium – krok pierwszy w pracy z AnnotationHelper

Przedmiotem pracy z AnnotationHelper jest zbiór, zawierające dane do adnotacji. Mogą to być dane tekstowe, wizualne (obrazki, a dokładniej URI obrazka), dźwiękowe (URI plików dźwiękowych (MP3 albo OGG), lub url streamingu z portalu soundcloud.com), video (pliki video (MP4, OGG), lub streaming do video z portali youtube albo vimeo albo dailymotion).

Przykładowe linki:

Obraz:

https://cn.plgrid.pl/repository/image/logo_PL_Grid-NG-45px.png

Muzyka:

https://w.soundcloud.com/player/?url=https%3A//api.soundcloud.com/tracks/71760739&auto_play=false&hide_related=false&show_comments=true&show_user=true&show_reposts=false&visual=true

Video:

<https://youtu.be/wygy721nzRc>, <https://vimeo.com/6864303>, <http://dai.ly/x4mliz>

Jak zdobyć linki z powyższych portali?

Soundcloud:

Pod wybranym utworem kliknij „Share”, potem „Embed”, skopiuj treść pola rozpoczynającego się od „<iframe”, wynij treść atrybutu *src*

YouTube:

Na stronie pożądanego filmu kliknij w „Udostępnij”, następnie jeszcze raz w „Udostępnij” poniżej pierwszego, skopiuj link

Vimeo:

Kliknij w „papierowy samolocik” pod ikoną serca i zegara po prawej stronie filmu. Skopiuj link z okienka, które się pojawiło.

Dailymotion:

Poniżej filmu kliknij w „Eksport” i skopiuj Permalink.

Możliwe jest wybranie dowolnej kombinacji powyższych mediów. Można poddawać adnotacji np. muzyki i tekstu, obrazu i muzyki, tekstu i video. Można również poddać adnotacji trzy i cztery z powyższych elementy na raz.

Wspierane formaty i struktury plików danych.

AnnotationHelper jest kompatybilny z dwoma formatami plików: CSV i JSON.

CSV

W przypadku wgrania do AnnotationHelper zbioru danych w formacie CSV obligatoryjny jest nagłówek, a w nim przynajmniej jedna z nazw kolumn *text*, *img_url*, *music_url*, *video_url*, przy czym wielkość znaków nie gra roli. **Wszystkie inne nazwy kolumn zostaną zignorowane!**

Kolumny te jest traktowana jako źródło danych, które mogą być poddane adnotacji.

Domyślnym separatorem jest średnik „;”, inne możliwe to: przecinek, znak tabulacji (TAB), spacja. Teksty, które w swojej treści posiadają znaki będące również separatorem (np. średnik) muszą być objęte cudzysłowami. Najlepiej, aby każdy tekst był objęty cudzysłowami.

AnnotationHelper zwróci błąd, jeśli:

- Zawartość zbioru nie jest kompatybilnym plikiem CSV
 - Brak nagłówka,
 - Różna liczba kolumn w poszczególnych wierszach
 - Teksty zawierające znak będący separatorem nie oznaczone dodatkowymi cudzysłowami.
 - Nazwy kolumn inne niż: *text*, *img_url*, *music_url*, *video_url*
 - Zawartość pól *img_url*, *music_url*, *video_url* niebędąca poprawnym URI
 - Zbiór danych zawiera więcej niż 10000 porcji danych
 - Zbiór zawiera znaki, które uniemożliwiają wczytanie go do natywnych bibliotek języka ruby. Pliki powinny być zgodne z kodowaniem UTF-8.

Przykładowy zbiór CSV:

Attr1;text;Attr3;video_url

cokolwiek;"Krótki tekst objęty cudzysłowiem ; średnik tu nic nie psuje";cokolwiek2;<http://dai.ly/x4mliz>

cokolwiek3;Dwa słowa;cokolwiek4;<https://vimeo.com/6864303>

więcej o csv na: https://pl.wikipedia.org/wiki/CSV_%28format_pliku%29

JSON

W przypadku zbioru danych w formacie JSON zakłada się, że zbiór to tablica zawierająca obiekty, z których każdy zawiera atrybut co najmniej jeden z atrybutów: *text*, *img_url*, *music_url*, *video_url*, (**Wszystkie inne zostaną zignorowane**), przy czym ostatnie trzy mogą wystąpić w jednym z dwóch miejsc w obiekcie:

1. bezpośrednio jako atrybut obiektu: [{"text": "Podдай mnie adnotacji", "img_url": "<http://link.do.obrazka>" }]
2. Jako atrybut obiektu nazwanego „other_details”, zgodnie ze schematem obowiązującym w usłudze syndykacji treści web 2.0, będącej częścią gridu Complex Networks. [{"text": "Podдай mnie adnotacji", "other_details": { "img_url": "<http://link.do.obrazka>" } }]

Uwagi dotyczące opisu:

1. W tablicy może się znajdować maksymalnie 10000 obiektów do adnotacji
2. Obiekty znajdujące się w tablicy mogą zawierać inne atrybuty. Nie zostaną one utracone w procesie adnotacji, zostaną one rozszerzone o dodatkowy atrybut zawierający wyniki adnotacji.

AnnotationHelper zwróci błąd, jeśli:

- Zawartość zbioru nie jest kompatybilnym plikiem JSON
 - Złe formatowanie
 - Zbiór danych zawiera więcej niż 10000 obiektów danych
 - Nazwy kolumn inne niż: *text*, *img_url*, *music_url*, *video_url*
 - Zbiór zawiera znaki, które uniemożliwiają wczytanie go do natywnych bibliotek języka ruby. Pliki powinny być zgodne z kodowaniem UTF-8.

Przykładowy zbiór w formacie JSON:

```
[{"text": "To jest przykładowy tekst", "video_url": "", "inny_atrybut": "wartość atrybutu"}, {"inny_atrybut": "wartość", "text": "To jest drugi tekst", "video_url": "", "inny_atrybut": "wartość atrybutu"}]
```

Więcej o json na: <https://pl.wikipedia.org/wiki/JSON>

Zdefiniowanie nowego projektu

W zakładce „Zdefiniuj nowy projekt” dostępny jest formularz konfigurowania nowych projektów. Link do tego formularza można również znaleźć w menu AnnotationHelper po zalogowaniu.

Dokładny opis poszczególnych parametrów został zamieszczony w dalszej części tego podręcznika oraz przy tych parametrach na stronie internetowej. Aby przeczytać opis, należy najechać myszą na ikonę z literką i (jak informacja), a wtedy wyświetli się chmurka (ang. *tooltip*).

W czasie definiowania projektu można w dowolnej chwili zapisać postęp klikając w guzik „zapisz”. W dowolnej chwili można edytować nierozpoczęty projekt wybierając go z listy dostępnej pod „Przeglądaj projekty” w menu. Aby rozpocząć badanie wciśnij „Zapisz i uruchom” z poziomu konkretnego formularza.

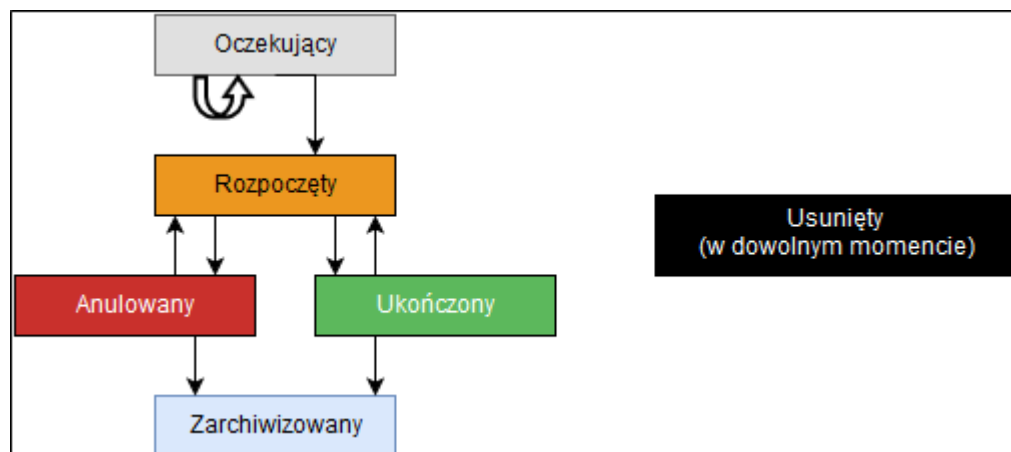
Każdy projekt w AnnotationHelper może posiadać jeden z czterech statusów: Oczekujący, Rozpoczęty, Ukończony, Anulowany. Możliwe jest jeszcze usunięcie

Statusy projektów i przyporządkowane im opcje

Każdy projekt ma w danej chwili jeden z następujących statusów:

- Oczekujący
- Rozpoczęty
- Ukończony
- Zarchiwizowany
- Anulowany
- Usunięty

Na rysunku poniżej pokazano jak projekty/badania mogą zmieniać status.



„Oczekujący”

Jest to projekt, który jest w trakcie konfigurowania. Projekt ma status oczekujący po tym jak zleceniodawca poda przynajmniej nazwę projektu i wciśnie guzik „Zapisz Konfigurację”.

Dostępne opcje w statusie „oczekujący”

Dostępne są trzy opcje dla projektów oczekujących: można podejrzeć ich szczegóły, edytować i usunąć.

„Rozpoczęty”

Projekt rozpoczęty to taki, który został zatwierdzony przez zleceniodawcę poprzez kliknięcie w guzik „Zatwierdź i Uruchom”, konfiguracja przeszła pomyślnie walidację, pożądaný zbiór danych został pomyślnie wczytany do systemu z repozytorium. Od tej chwili zleceniodawca ma dostęp do specjalnych linków umożliwiających dostęp do adnotacji w ramach tego projektu. Linki te może rozesłać do pożądaných osób. Jeśli zleceniodawca podczas konfiguracji oznaczył projekt jako publiczny, to od tej chwili projekt będzie również widoczny dla wszystkich (również dla osób niezarejestrowanych) w zakładce „Publiczne projekty”.

Dostępne opcje w statusie „rozpoczęty”

Poza szczegółami, edycją i usunięciem, zleceniodawca projektu może również:

1. Podglądać aktywność osób adnotujących,
2. skopiować wspomniany link dostępowy,

3. pobrać próbkę danych w tym samym formacie co ostateczny zbiór, który będzie dostępny po zakończeniu badania (np. do wstępnych testów programów, które zlecniodawca ma zamiar użyć zbiór z adnotacjami)
4. anulować projekt.

Jeśli projekt osiągnął kilka procent ukończenia, wtedy pojawia się pasek postępu oraz szacowany termin ukończenia.

Edycja rozpoczętego projektu

System AnnotationHelper pozwoli na edycję tylko niektórych parametrów konfiguracyjnych badania:

- słowny opis badania,
- treść polecenia
- skróty klawiszowe etykiet
- sposób wybierania kolejności treści do adnotacji
- ręczne dodanie kont adnotatorów i (automatyczne) wygenerowanie dla nich linków dostępowych.

„Anulowany”

Projekt staje się anulowanym po kliknięciu w opcję „anuluj projekt” oraz potwierdzeniu tej chęci za pośrednictwem okienka, które się pojawi. Od tej chwili adnotujący stracą dostęp do projektu oraz pojawią się nowe opcje w pasku opcji.

Anulowanie projektu to akcja, którą wybierze zlecniodawca widzący, że poziom adnotacji jest wg niego/niej wystarczający lub aktywność adnotatorów zanikła, co nie daje szans na dokończenie adnotacji. Po anulowaniu adnotacji istnieje możliwość pobrania zaadnotowanej części zbioru.

Dostępne opcje w statusie „anulowany”

W tym statusie pojawiają się dwie nowe opcje: Wygeneruj pliki danych oraz Wznów projekt. Wznowienie oczywiście przeniesie projekt w stan „rozpoczęty”. Wygenerowanie plików działa tutaj tak samo jak przy statusie „ukończony”. Opcja ta jest możliwa tylko gdy projekt przekroczy 10% ukończenia.

„Ukończony”

Projekt nabiera ten status po uzyskaniu odpowiedniej liczby adnotacji zależnej od parametru [Liczba adnotacji pojedynczego obiektu danych](#) oraz wielkości zbioru. Istnieje możliwość powrotu do statusu rozpoczęty. Stanie się to wtedy, kiedy zlecniodawca z poziomu podglądu adnotatorów stwierdzi wyraźną fałszywość adnotacji jednego z nich i postanowi skasować je z projektu. Wtedy ktoś inny będzie miał szansę wprowadzenia brakujących adnotacji.

Dostępne opcje w statusie „ukończony”

Projekt w statusie ukończony ma podobne opcje jak w przypadku projektów anulowanych. Po wygenerowaniu plików danych w miejsce opcji „wygeneruj...” pojawia się opcja pobrania danych.

Opcja „Wygeneruj pliki danych”

Wygenerowanie plików danych polega na agregacji odpowiedzi adnotatorów, utworzeniu pliku z tych danych, wgranie pliku do repozytorium jako kolejny plik w tzw. Itemie repozytorium dspace.

„Zarchiwizowany”

Projekt przechodzi w stan zarchiwizowany po wybraniu takiej opcji przez użytkownika, lub po interwencji administratora. Archiwizowanie zbiorów danych pozwala na szybsze działanie systemu, dlatego zaleca się archiwizować swoje projekty po ich zakończeniu.

Dostępne opcje w statusie „zarchiwizowany”

Po archiwizacji projekt traci większość opcji. Można wciąż usunąć całą treść oraz pobrać wynikowy plik z danymi. Dane pobierane są z repozytorium za pośrednictwem AnnotationHelper.

Opis poszczególnych parametrów projektów

Nazwa projektu

Nazwy projektów pozwalają w łatwy sposób odróżnić jeden od drugiego. W przypadku projektów publicznych pomagają zainteresować odwiedzających stronę do poświęcenia swojego czasu na adnotację danego projektu.

Opis projektu

Podobnie jak nazwa, ale pozwala na zapis dłuższej treści.

Treść polecenia dla adnotujących

Domyślne polecenia jakie zobaczy adnotator to: „Dokonaj adnotacji treści korzystając z poniższych etykiet!”. Jeśli życzysz sobie, możesz wpisać inną treść.

Wybór zbioru danych

Z repozytorium

AnnotationHelper wyświetli wszystkie zbiory dodane do repozytorium z zaznaczeniem opcji oznaczającej chęć użycia zbioru w tym systemie. Z wyświetlonych można wybrać dokładnie jeden. Repozytorium umożliwia dodanie kilku plików na raz, ale adnotacji będzie poddany tylko jeden-pierwszy dodany.

Z pliku na własnym dysku

Istnieje możliwość wgrania swojego zbioru bezpośrednio z dysku. Akceptowane formaty plików to: *.json, *.csv. Po wybraniu zbioru należy wypełnić krótki formularz dot. danych opisujących zbiór danych.

Format danych

Zleceńodawca może pomóc systemowi w detekcji formatu danych oznaczając wprost, czy jest to JSON, czy CSV.

Określenie co podlega adnotacji

Adnotacji może podlegać tekst, obrazki, dźwięki/muzyka i film oraz dowolne kombinacje powyższych.

Liczba etykiet

Zaznacz, spośród jakiej liczby etykiet adnotatorzy będą wybierać.

Tekst etykiet i skróty klawiszowe

Etykiety

Ten parametr należy do najważniejszych. Wpisuje się tu etykiety, spośród których wybierać będą adnotatorzy.

Skróty klawiszowe

W celu ułatwienia adnotacji, zwłaszcza w przypadku małej liczby etykiet (np. 3), warto jest zdefiniować skróty klawiszowe, które zaznaczą daną etykietę. Kontrolka skrótu klawiszowego działa inaczej niż pozostałe, gdyż wyświetlać ona będzie nie tradycyjny tekst, tylko kombinację wciśniętych klawiszy, np. zamiast wykrzyknika '!' zobaczymy shift+1. Nie wszystkie skróty są dozwolone. Jeśli po wciśnięciu klawiszy wpis w kontrolce się nie zmienił, to jest to znak, że dany skrót jest niedozwolony.

Alternatywny sposób wpisywania etykiet

W przypadku większej liczby etykiet wygodniejszym sposobem podania ich jest kliknięcie „Wprowadź...”, co spowoduje otwarcie osobnego okna, w które można wkleić etykiety jedna pod drugą. Po wciśnięciu klawisza "Zatwierdź" okno się zamknie, a etykiety zostaną wpisane w odpowiednie pola.

Adnotacja jedno- lub wieloetykietowa

Przy adnotacji jednoetykietowej każdemu obiektowi danych przypisywana jest dokładnie jedna etykieta ze wszystkich możliwych.

Przy adnotacji wieloetykietowej możliwa jest dowolna kombinacja zaznaczeń etykiet.

Liczba adnotacji pojedynczego obiektu danych

Dobłą praktyką ręcznego etykietowania / adnotowania danych jest adnotacja każdego obiektu danych przez więcej niż jedną osobę. Wybierając wartość tego parametru na przykład 3 spowoduje, że do ukończenia projektu każdy obiekt danych będzie musiał być zaadnotowany 3 razy.

Sposób wyboru kolejności obiektów ze zbioru

AnnotationHelper pozwala na wybór trzech sposobów wybierania danych do adnotacji:

1. wg kolejności od najrzadziej adnotowanych
2. wg kolejności od najczęściej adnotowanych
3. wg kolejności wystąpienia w zbiorze

Wybranie pierwszej opcji spowoduje równomierną adnotację wszystkich obiektów. Wybranie drugiej spowoduje szybkie osiągnięcie minimalnej liczby adnotacji poszczególnych obiektów. Wybranie trzeciej opcji da zbliżony efekt jak w drugiej. Różnice pojawiają się głównie w sytuacjach kasowania adnotacji podejrzanych adnotatorów przez zleceniodawcę.

Poziom dostępności projektu

W AnnotationHelper wyróżnia się trzy poziomy dostępności projektu:

1. Publiczny
2. Prywatny
3. Chroniony

Do publicznych projektów dostęp ma każdy (np. poprzez portal AnnotationHelper, linki dostępne). Do projektów prywatnych dostęp mają osoby posiadające link dostępowy publiczny i adnotatorzy zdefiniowani powyżej posiadający link dostępowy prywatny. Do projektów chronionych dostęp mają tylko powyżsi adnotatorzy. Zleceniodawca projektu (TY) ma zawsze dostęp do adnotacji swoich projektów.

Linki dostępne i podgląd postępów

W zależności od wybranego poziomu dostępności linki dostępne będą bardziej lub mniej ważne dla projektu.

	Drogi dostępu do interfejsu adnotacji		
Rodzaj projektu	Strona internetowa	Link publiczny	Link dostępowy prywatny (imienny)
Publiczny			
Prywatny	x		
Chroniony	x	x	

Uwaga! W każdym przypadku zleceniodawca projektu ma dostęp do interfejsu adnotacji.

Gdzie znajdę linki dostępne

Link publiczny

Projekty o statusie „rozpoczęty” oraz poziomie dostępności publicznym lub prywatnym mają zamieszczony link publiczny pod ikoną "Link" na liście opcji. Można kliknąć w nią prawym przyciskiem myszy i wybrać opcję „kopiuj adres linku” lub zbliżoną do tej (dokładna treść zależy od przeglądarki).

Link ten również można znaleźć pod opcją „Szczegóły projektu”

Link prywatny (imienny)

Linki te są generowane przez system dla każdego zdefiniowanego przez zleceniodawcę adnotatora. Można je uzyskać z zakładki "Adnotatorzy" z menu projektu.

Kontrola nad użytkownikami

Zleceniodawca uzyskuje wsparcie ze strony systemu również w kontroli nad osobami adnotującymi. W zakładce "Adnotatorzy" dostępne są opcje: Podgląd Adnotacji, Zablokowanie dostępu, Usunięcie adnotacji "tego" użytkownika pod każdym adnotującym. Podgląd adnotacji osoby pozwala zobaczyć losowo wybrane adnotacje w celu oceny ich poprawności. Jest to mechanizm obronny przed sytuacjami takimi, że osoba adnotująca celowo wprowadza fałszywe adnotacje. Po wykryciu takiej osoby zleceniodawca może zablokować dla niej dostęp do projektu oraz usunąć wszystkie dotychczasowe adnotacje. Raz zablokowany dostęp można później przywrócić.

Podręcznik dla osoby adnotującej

Również osoby adnotujące mają do swojej dyspozycji podręcznik. Jest on generowany dynamicznie na podstawie konfiguracji projektu. Jest on dołączony do każdego projektu i rozwinięty przy pierwszym załadowaniu strony. Podręcznik można przejrzeć a potem zwinąć/ukryć.