Dokumentacja AF-Words-Clustering

Spis treści

[Wprowadzenie 1](#_Toc493584264)

[Wymagania systemowe 1](#_Toc493584265)

[Opis techniczny 1](#_Toc493584266)

[Instrukcja użytkowania 2](#_Toc493584267)

## Wprowadzenie

Program AF-Words-Clustering to program konsolowy napisany w języku Python mający na celu grupowanie w klastry podanych słów wejściowych. Program grupuje słowa za pomocą algorytmu [Affinity Propagation Clustering](https://www.wikiwand.com/en/Affinity_propagation). Podane słowa wejściowe przed grupowaniem są sprowadzane do swej formy podstawowej (jeśli takowa istnieje) za pomocą algorytmu stem [Porter2](http://snowball.tartarus.org/algorithms/porter/stemmer.html).

## Wymagania systemowe

* Windows 7, Windows 8.1, Windows 10
* Python 3.5.2 lub nowszy.

## Opis techniczny

Program napisany jest w języku Python 3 korzystając z bibliotek distance, stemming, scikit-learn oraz numpy. Za pomocą pakietu stemming słowa wejściowe są sprowadzane do formy podstawowej, usuwane są także białe znaki oraz duplikaty. Na podstawie tak wstępnie przetworzonych danych generowana jest macierz podobieństw (każdego słowa z każdym). Do tego celu z pakietu distance użyty jest dystans [Levenshteina](https://www.wikiwand.com/pl/Odleg%C5%82o%C5%9B%C4%87_Levenshteina) pomiędzy słowami. Wspomniana macierz stanowi bezpośredni parametr wejściowy dla algorytmu Affinity Propagation Clustering. Słowa wejściowe wraz z przyporządkowaniem do grupy są zwracane w formie pliku CSV.

Szczegółowa dokumentacja techniczna w języku angielskim dostępna jest w archiwum **af\_words\_clustering\_technical\_doc.zip**.

## Instrukcja użytkowania

Program dystrybuowany jest jako pakiet Python EGG dzięki czemu może być używany jako program wykonywalny jak i łatwo zainstalowany jako biblioteka. Plik EGG jest archiwum ZIP dzięki czemu wewnątrz niego dołączone zostały przykłady pliku wejściowego. Przykładowy plik - example.txt można znaleźć w folderze example-dataset (1000 słów angielskich).

Pierwsze uruchomienie:

1. Instalacja jako biblioteka, instalacja zależności:  
   *easy\_install af\_words\_clustering-0.1.0-py3.5.egg*
2. Uruchomienie programu:   
   *python af\_words\_clustering-0.1.0-py3.5.egg -i C:\example-dataset\example.txt -o C:\example-dataset\example-output.txt -s False -a dynamic*

Parametry:

* -i <ścieżka pliku wejściowego>
* -o <ścieżka pliku wyjściowego>
* -s – Flaga logiczna. Jeśli *True* program będzie liczyć współczynnik Silhouette. W przypadku *False* krok ten zostanie pominięty.
* -a – [Preferencja punktów](http://www.psi.toronto.edu/affinitypropagation/faq.html). (Zmniejszenie wartości sprzyja zmniejszeniu ilości klastrów)
  + *auto* – jako preferencja zostanie przyjęta mediana wszystkich punktów
  + *dynamic* – jako preferencja zostanie przyjęta minimalna wartość podobieństwa w macierzy podobieństw
  + <integer> np. 300 – arbitralnie podana wartość preferencji