

# 1 Uruchomienie programu

Z programu korzystamy za pomocą wiersza poleceń. Aby uruchomić wiersz poleceń wybieramy:

```
Start→Uruchom→cmd
```

Następnie przechodzimy do katalogu z programem:

```
cd C:\gdziesnadysku\rainbow-win\
```

Alternatywnie można otworzyć katalog z programem (rainbow-win) w trybie graficznym i przytrzymując Shift kliknąć prawy przycisk myszy. Rozwinie się menu z którego należy wybrać „otwórz okno polecenia tutaj”.

## 2 Pierwsze kroki

### 2.1 Zdefiniowanie zmiennej środowiskowej home

Przy pierwszym uruchomieniu programu należy zdefiniować zmienną środowiskową home:

```
set HOME=.
```

### 2.2 Zapoznanie się z dostępnymi opcjami programu

Przed rozpoczęciem pracy z programem warto zapoznać się z opcjami kontrolującymi jego działanie. Program akceptuje około 160 opcji, jednak w większości przypadków program działa skutecznie z domyślnymi wartościami opcji, co ułatwia jego obsługę.

Do wygenerowania opisu opcji służy przełącznik --help, do wygenerowania wyłącznie nazw opcji --usages. W celu zapoznania się z opisami opcji warto przekierować rezultat polecenia do pliku:

```
rainbow --help > pomoc.txt
```

## 3 Dane uczące

W najbardziej podstawowym przypadku dane uczące powinny być w zwykłych plikach tekstowych (jeden dokument tekstowy - jeden plik). Pliki powinny być umieszczone w folderach, w taki sposób, aby dokumenty które zostały zaklasyfikowane do jednej klasy decyzyjnej znajdowały się w jednym folderze.

W ćwiczeniu skorzystamy z ze zbioru 20\_newsgroups. Jest to zbiór 20000 wiadomości poczty elektronicznej, każda z wiadomości należy do jednej z 20 grup dyskusyjnych. Pliki umieszczone są w następujących katalogach:

```
comp.graphics  
comp.os.ms-windows.misc  
comp.sys.ibm.pc.hardware  
comp.sys.mac.hardware  
comp.windows.x  
rec.autos
```

*rec.motorcycles*  
*rec.sport.baseball*  
*rec.sport.hockey*  
*sci.crypt*  
*sci.electronics*  
*sci.med*  
*sci.space*  
*misc.forsale*  
*talk.politics.misc*  
*talk.politics.guns*  
*talk.politics.mideast*  
*talk.religion.misc*  
*alt.atheism*  
*soc.religion.christian*

Poszczególne katalogi to nazwy grup dyskusyjnych. Mogą one być traktowane jako klasy dokumentów.

## 4 Klasyfikacja tekstów

Klasyfikacja tekstów odbywa się w dwóch etapach. W pierwszym etapie program uruchamiany jest w trybie indeksującym - na podstawie obliczeń tworzona jest reprezentacja numeryczna podanych plików wejściowych, która zapisywana jest w postaci modelu. Po stworzeniu modelu można na jego podstawie klasyfikować nowe teksty.

### 4.1 Indeksowanie

W trybie indeksującym określamy katalog w jakim ma znaleźć się model za pomocą przełącznika `--data-dir nazwa_katalogu` (`'-d nazwa_katalogu'`). Jeśli nazwa katalogu nie zostanie podana program użyje katalogu `~/rainbow`. W przypadku kiedy podany katalog nie istnieje, zostanie on stworzony przez program. Dane do zaindeksowania należy podać jako sekwencję nazw katalogów. Nazwy katalogów zostaną utożsamione z nazwami grup. Przykładowo, jeśli chcielibyśmy stworzyć model dla klas (grup):

```
talk.politics.misc talk.politics.guns talk.politics.mideast
```

w katalogu model skorzystamy z następującego polecenia:

```
rainbow -d model 20newsgroups/talk.politics.misc 20newsgroups/talk.politics.guns  
20newsgroups/talk.politics.mideast
```

W rezultacie wygenerowane zostaną reprezentacje danych które posłużą do klasyfikacji nowych dokumentów do trzech wymienionych klas. W przypadku większej liczby klas wpisywanie sekwencji nazw może być uciążliwe, zamiast tego możemy skorzystać z ze znaków specjalnych: `"*" i "?"`. Korzystając ze znaku `"*"` możemy zapisać powyższe polecenie jako:

```
rainbow -d model --index 20newsgroups/talk.politics.*
```

Pliki ze zbioru 20newsgroups posiadają w nagłówku nazwę grupy dyskusyjnej. Nie chcemy tego ujmować w modelu w związku przy tworzeniu modelu należy posłużyć się opcją

--skip-header:

```
rainbow -d model --index --skip-header 20newsgroups/talk.politics.*
```

## 4.2 Informacje na temat modelu

Po wykonaniu indeksowania możemy uzyskać informacje na temat stworzonego modelu, np.

Wartość informacyjną słów (5 najlepszych):

```
rainbow -d model --print-word-Infogain 5
```

Prawdopodobieństwa pojawienia się słów w danej klasie :

```
rainbow -d model -T 5 --print-word-probabilities=talk.politics.mideast
```

Prawdopodobieństwo pojawienia się danego słowa w klasach:

```
rainbow -d model --print-word-counts=proper
```

Można również wyświetlić macierz modelu, macierz zawiera nazwę dokumentu i liczbę określającą ile razy dane słowo pojawia się w tym dokumencie. Jedna linia odpowiada jednemu dokumentowi:

- pierwsza pozycja to ścieżka do dokumentu
- druga pozycja to klasa do jakiej został zaklasyfikowany tekst
- kolejne pozycje to słowa w dokumencie wraz z liczbą wystąpień słowa w dokumencie

Aby wyświetlić macierz korzystamy z następującego polecenia:

```
rainbow -d model --print-matrix
```

Aby ograniczyć rozmiar wyniku polecenia możemy skorzystać z odpowiednich opcji programu. Przykładowo, aby wyświetlić pierwszych 10 wpisów macierzy ograniczonej do 100 słów, w której dla każdego dokumentu wypisane zostaną tylko słowa które się w nim pojawiły:

```
rainbow -d model -T 100 --print-matrix=siw | head -n 10
```

## 4.3 Klasyfikacja nowych tekstów

### 4.3.1 Klasyfikacja pojedynczego tekstu

Aby sklasyfikować pojedynczy tekst korzystamy z opcji '--query'. Wówczas podajemy tekst na wejście standardowe. Ciąg znaków należy zakończyć sekwencją znaków Ctrl-Z (Ctrl-D w systemie Unix). Jeśli zbudowano model w katalogu "model" można wywołać klasyfikator następującym poleceniem:

```
rainbow -d model --query
```

Po otrzymaniu komunikatu:

```
"Type your query text now. End with a Control-D."
```

można przystąpić do wprowadzania tekstu który ma być klasyfikowany, np.

```
"A imię jego czterdzieści i cztery  
Ctrl-Z"
```

Co oznacza wynik klasyfikacji jaki został uzyskany dla takiego tekstu? Jaki wynik uzyskamy wpisując "israeli, guns"?

### 4.3.2 Klasyfikacja wielu tekstów i ocena klasyfikacji

Klasyfikacji nowych plików można dokonać za pomocą przełącznika

```
'--test-files nazwa_katalogu'
```

Nowe pliki powinny znajdować się w katalogach o nazwach odpowiadających nazwom klas w modelu. Informacja ta nie jest brana pod uwagę przy klasyfikacji, jest natomiast wykorzystywana przy ocenie klasyfikacji. Klasyfikacja nowych plików nie musi być uprzednio znana, jednak ze względów technicznych nowe pliki należy umieścić w katalogach odpowiadających nazwom katalogów wykorzystanym podczas uczenia.

W przypadku oceny klasyfikacji z wykorzystaniem podziału na dane uczące i testujące stosowana jest metoda hold-out, która przyjmuje dwa parametry: liczbę plików które zostaną wylosowane i potraktowane jako zbiór testujący,  $i$  i liczbę powtórzeń losowania podziału na zbiór uczący i testujący. W przypadku podania liczby całkowitej jako liczby plików do zbioru testującego zostanie wylosowana taka właśnie liczba plików, w przypadku podania liczby rzeczywistej będzie to odpowiedni procent.

Poniższe polecenie:

```
rainbow -d model --test-set=0.5 --test=1
```

spowoduje jednokrotny podział na zbiór uczący i testujący, w którym połowa plików zostanie potraktowana jako zbiór uczący. Wynik działania polecenia to następująca linijka dla każdego pliku:

```
ścieżka_do_pliku rzeczywista_klasa przewidziana_klasa:punkty1 druga_przewidziana_klasa:punkty2  
...
```

Wygenerowane dane mogą zostać przetworzone przez narzędzia służące do przetwarzania tekstów, takie jak AWK czy Perl.