Zastosowanie biogramów liter w detekcji języka tekstu:

1) Ściągnij z Internetu i zapisz do plików po 5 przykładowych stron tekstu w językach:

- angielski (zapis o pliku ang.txt)
- polski (zapis o pliku pol.txt)
- niemiecki (zapis o pliku niem.txt)
- hiszpański (zapis o pliku his.txt)
- włoski (zapis o pliku wlo.txt)

2) Przekonwertuj ściągnięte strony tekstu na alfabet bez znaków specjalnych,

a b c d e f g h i j k l m n o p r s t u w x y z

np. ą na a, ę na e, ł na l, itd.,

3) Wygeneruj bigramy liter alfabetu z punktu 2,

czyli:

aa, ab, ac, ..., az

ba, bb, bc, ..., bz

...

za, zb,..., zz

4) Wylicz częstość pojawiania się wszystkich bigramów liter w pięciu plikach z punktu 1,

5) Sprawdź czy za pomocą bigramów liter możliwa jest detekcja języka tekstu.

6) Sprawdź czy detekcja języka może bazować na częstości występowania liter.