mj_lda_rf

Maciej Jankowski

20 marca 2017

Celem tego dokumentu jest przetestowanie metody LDA w dwóch wariantach

1. Przy użyciu unigramów
2. Przy użyciu unigramów i bigramów

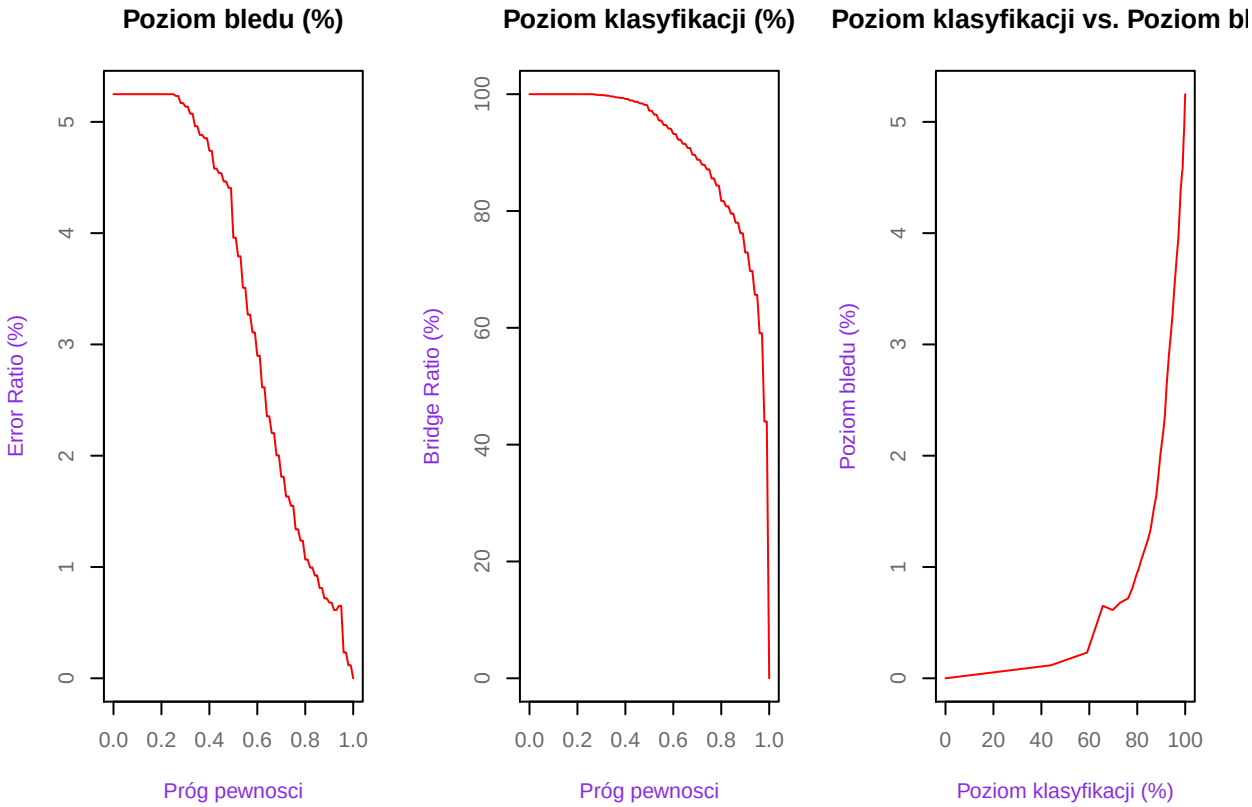
Model LDA zakłada, że ilość tematów K, jest z góry ustalona. Dlatego analizaę rozpoczniemy od oszacowania ilości tematów w naszym zbiorze danych. Rozpatrzymy cztery metryki:

1. Griffiths2004
2. CaoJuan2009
3. Arun2010
4. Deveaud2014

Obliczenia zostały wykonane przy użyciu biblioteki *ldatuning* zaimplementowanej w języku R. Następnie, rozpatrzymy klasyfikację dokumentów przy użyciu algorytmu random forest. Klasyfikacji dokonamy w zredukowanej przestrzeni, gdzie każdy dokument jest reprerzentowany przez wektor długości K. Poszczególne składowe tego wektora są prawdopodobieństwami tematu w dokumencie $\mathcal{P}(Z_k|\boldsymbol{d})$.

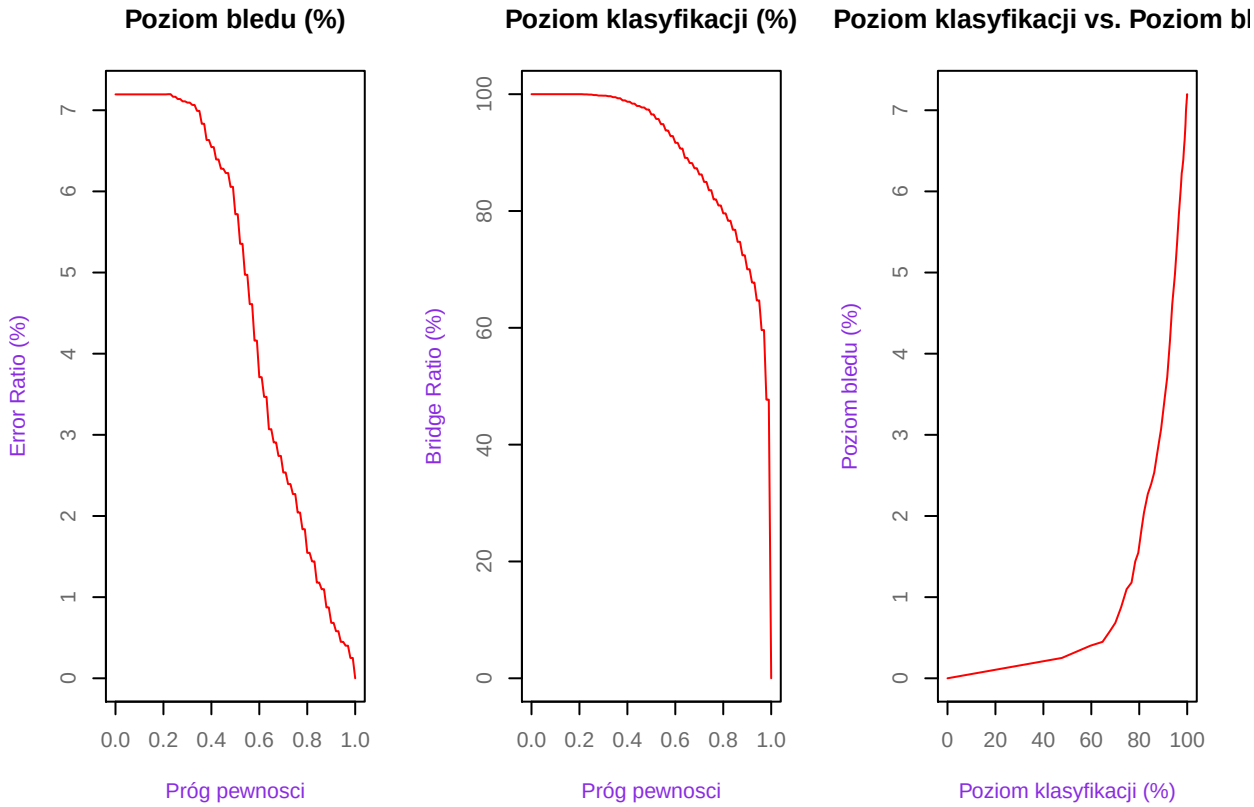
W pierwszym przykładzie zastosowaniśmy model LDA do danych. W pierwszym kroku stworzyliśmy tabelę TF, w której kolumnami były poszczególne termy. Wyniki zostały przedstawione na rysunkach

Unigrams



W kolejnym eksperymencie, zastosowaliśmy ten sam model. Użyliśmy jednak innej tabeli TF. Tym razem w kolumnach znalazły się zarówno unigramy jak i bigramy, czyli dwuwyrazowe frazy.

Bigrams



Wynik tego eksperymentu pokazuje, że model oparty o unigramy uzyskał lepszą jakość klasyfikacji. Przyczyną tego może być fakt, że użycie bigramów prowadzi do przeuczenia modelu. Aby tego uniknąć należało by wprowadzić jakiś współczynnik wygładzania np

$$\mathcal{P}(w_t|w_{t-1}) = \lambda\frac{N_i}{N} + (1-\lambda)\frac{N_{i|j}}{N_j},$$

gdzie N_i oznacza ilość wystąpień słowa w_i w korpusie, a $N_{i|j}$ oznacza ilość wystąpień słowa w_i bezpośrednio po słowie w_j.

W literaturze możemy znaleźć następujące rozszerzenia modelu LDA opartego o unigramy.

1. Topic Modeling: Beyond Bag-of-Words autorstwa Hanna M. Wallach
2. Improvements to the Bayesian Topic N-gram Models autorstwa Hiroshi Noji, Daichi Mochihashi, Yusuke Miyao

Dalsze pomysły

1. Redukcja wymiarów oparta o kodowanie arytmetyczne
2. Kodowanie arytmetyczne wykorzystujące rozkłady w tematach i rozkłady tematów
3. Rozszerzyć modelowanie tematyczne o drugie kryterium optymalizacji - minimalizacja entropii w ramach tematu. Chodzi o to, żeby tematy były jak najbardziej specyficzne.