Automatyczna klasyfikacja i ekstrakcja tematu krótkich wiadomości tekstowych w języku polskim

Paweł Obrok
promotor: dr. Michał Korzycki

Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie
Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki
Katedra Informatyki

25 października 2012

Latent Semantic Indexing

Metoda oparta o algebraiczną redukcję rzędu macierzy

Latent Semantic Indexing

Metoda oparta o algebraiczną redukcję rzędu macierzy

Istota: znaleźć macierz o zadanym rzędzie, która najlepiej przybliża daną macierz w sensie normy Frobeniusa

Latent Dirichlet Allocation

Metoda oparta o model generatywny dokumentów

Latent Dirichlet Allocation

Metoda oparta o model generatywny dokumentów

- Wybierz dystrybucję tematów z $D(\alpha)$

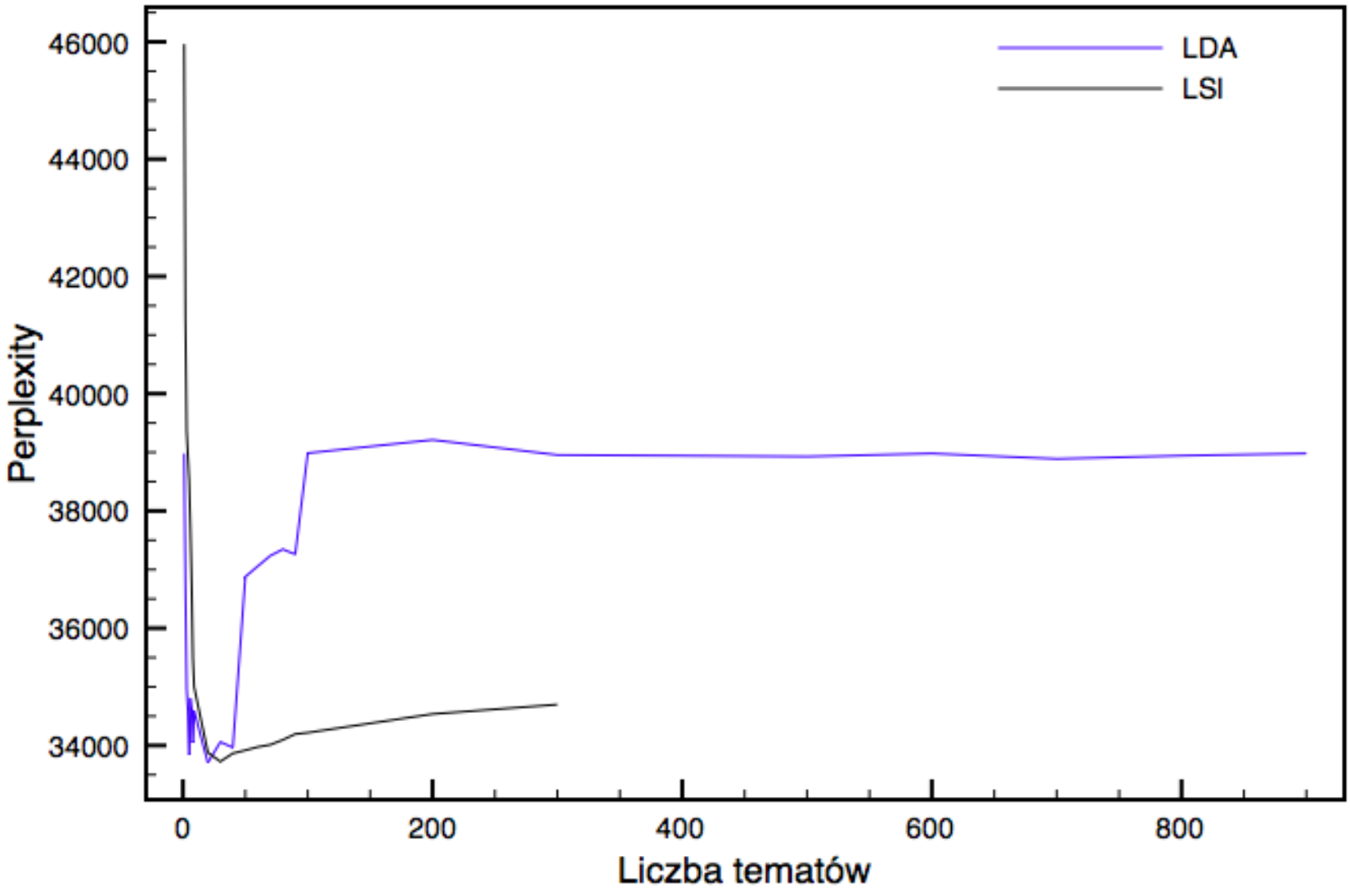
Latent Dirichlet Allocation

Metoda oparta o model generatywny dokumentów

- Wybierz dystrybucję tematów z $D(\alpha)$
- Powtarzaj
  - Wybierz temat z dystrybucji tematów
  - Wybierz wyraz z dystrybucji zadanej przez temat

Perplexity

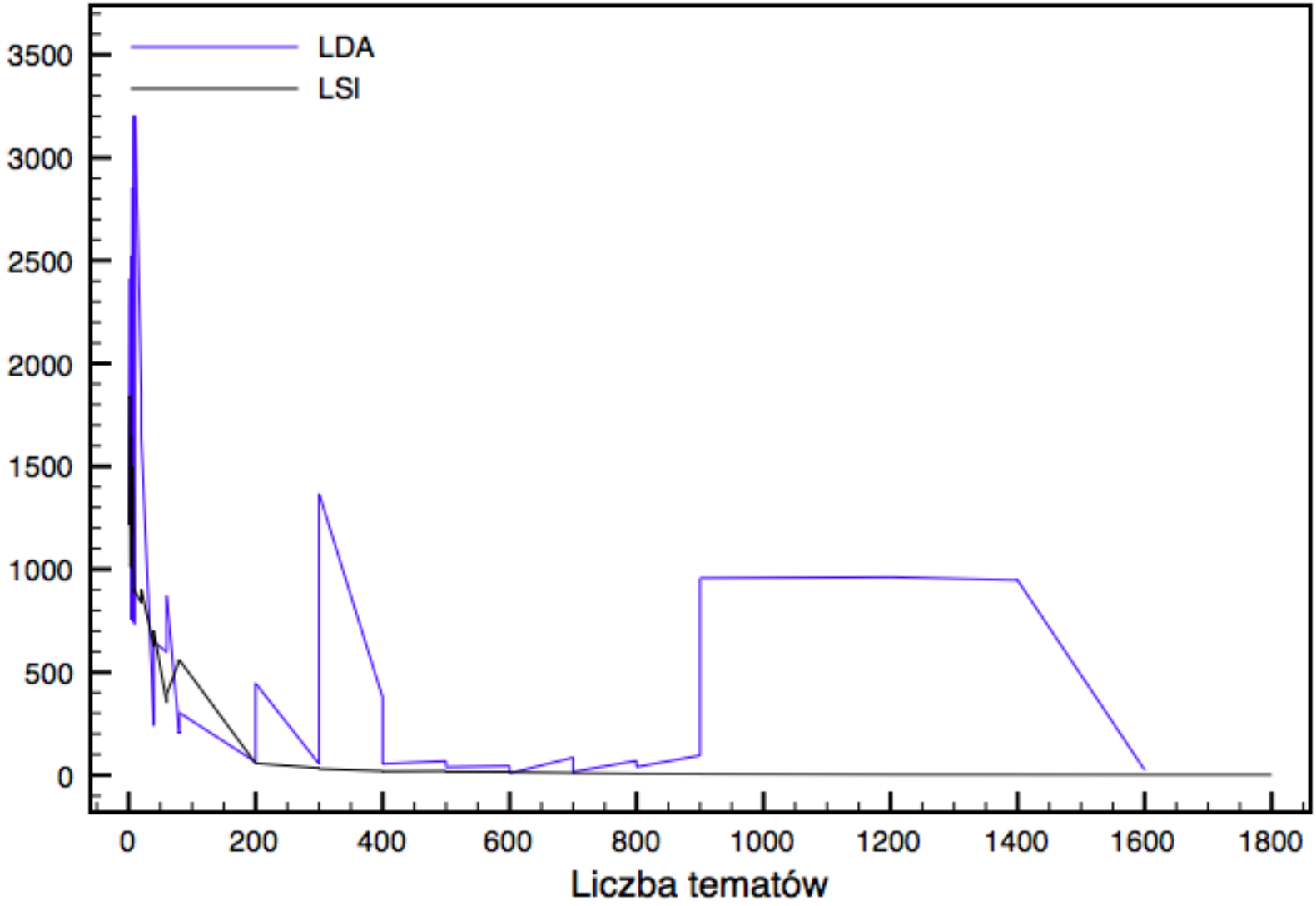
Perplexity


LDA
LSI
Perplexity
46000
44000
42000
40000
38000
36000
34000
0
200
400
600
800
Liczba tematów

Przykładowy problem

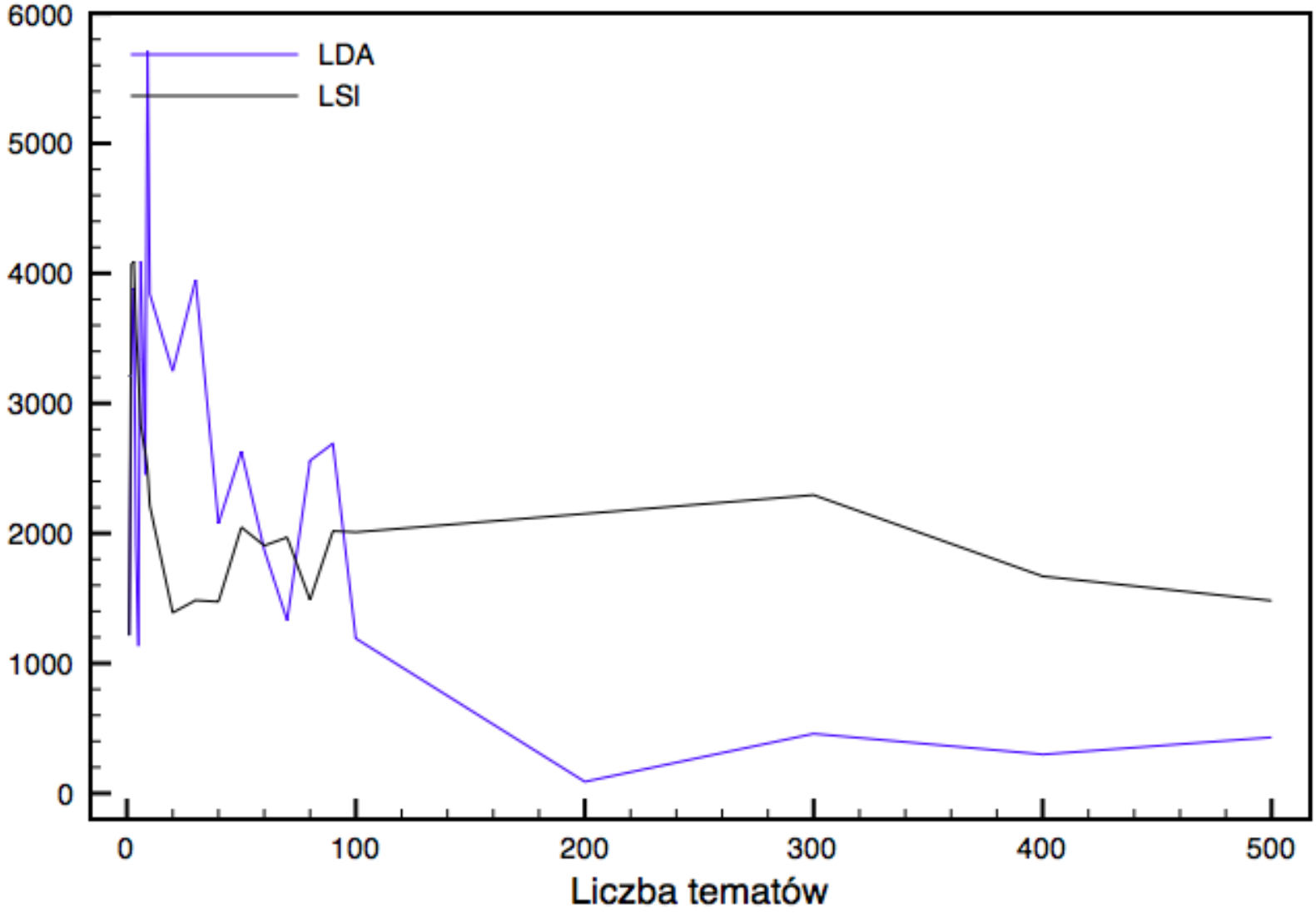
Przykładowy problem

Suma kwadratów ranków skojarzonych dokumentów w zwróconych wynikach z wykorzystaniem CLP



Przykładowy problem

Suma kwadratów ranków skojarzonych dokumentów w zwróconych wynikach bez wykorzystania CLP



Wnioski

Wnioski

- LDA spisuje się gorzej niż LSI

Wnioski

- LDA spisuje się gorzej niż LSI
- LDA jest mniej stabilne niż LSI

Wnioski

- LDA spisuje się gorzej niż LSI
- LDA jest mniej stabilne niż LSI
- LDA lepiej znosi większy rozmiar słownika

Wnioski

- LDA spisuje się gorzej niż LSI
- LDA jest mniej stabilne niż LSI
- LDA lepiej znosi większy rozmiar słownika
- LDA lepiej skaluje się z rozmiarami korpusu