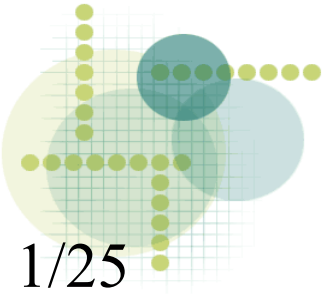
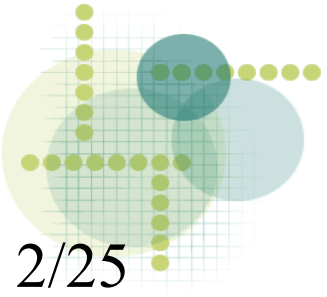
Przewidywanie cen akcji z wykorzystaniem artykułów prasowych

Mateusz Kobos, 05.12.2007
Seminarium Metody Inteligencji Obliczeniowej



Spis treści

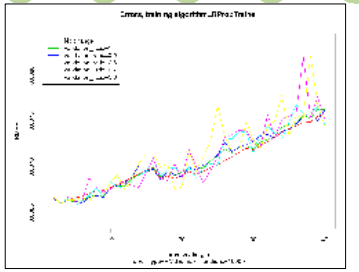


Ogólna budowa programu

ściąganie, parsowanie, czyszczenie

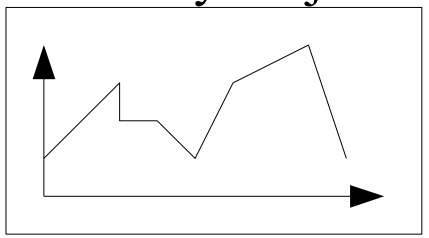
BD

wyniki



artykuły

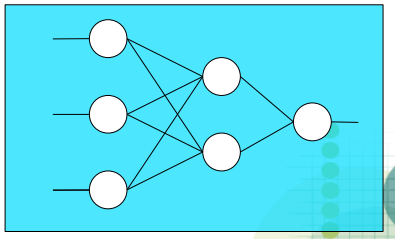
ceny akcji



WWW

analiza wyników

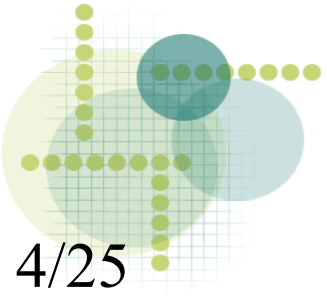
preprocessing, zamiana na liczby



model

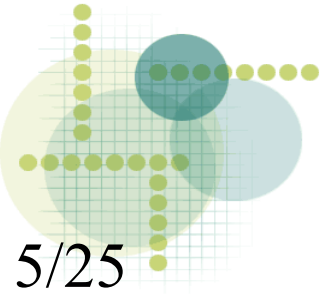
Ogólna budowa programu

- Główne wykorzystane technologie
  - MySQL
  - Python (szkielet programu, większość kodu)
    - Porter stemmer
  - C++ (algorytmy wymagające wydajności: model (ANN))
    - biblioteka FANN



Pobieranie dostępnych danych

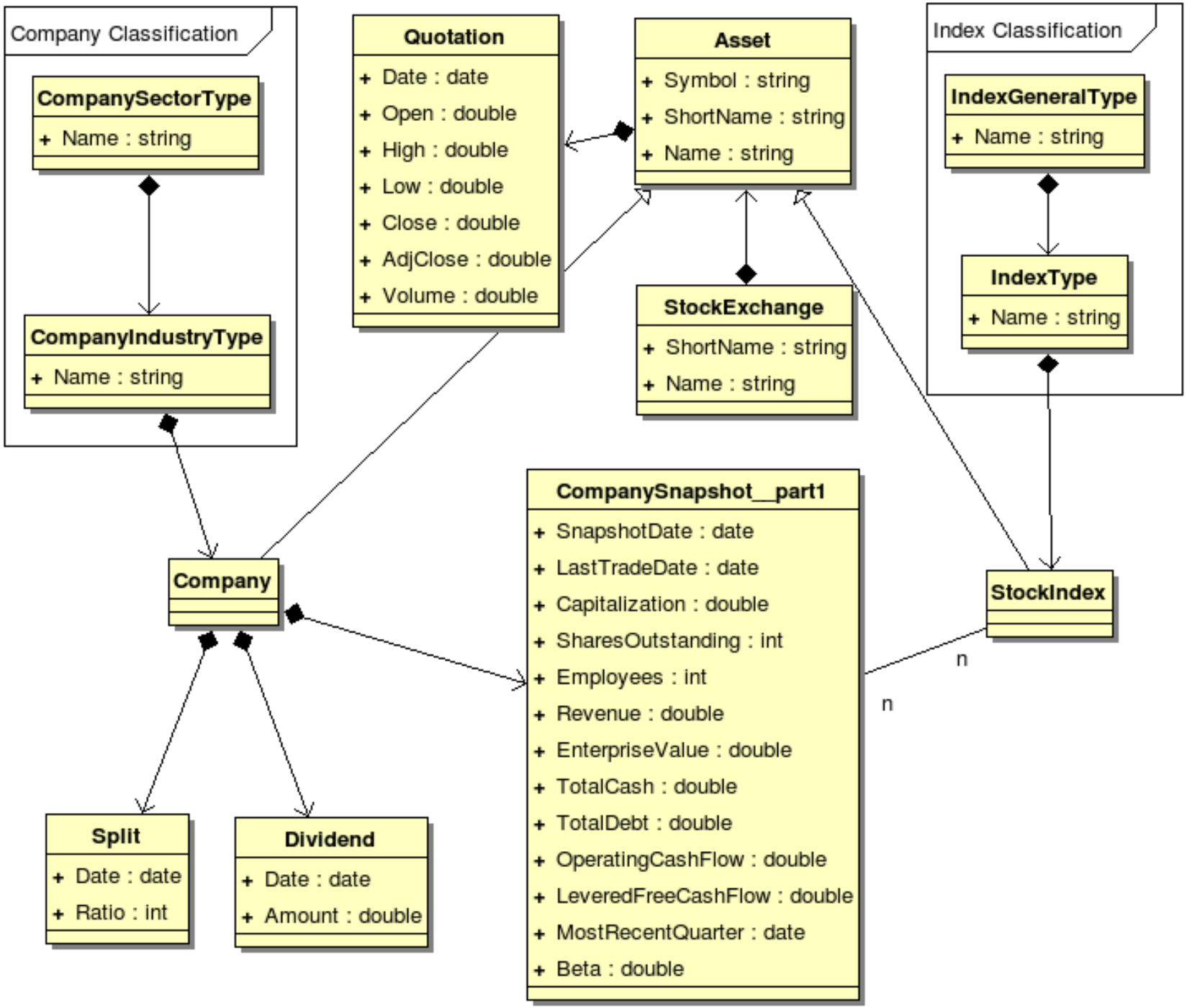
- Notowania akcji i indeksów
  - Yahoo.com
  - dzienne notowania znormalizowane ze względu na split-y i dywidendy
- Artykuły
  - Proquest
  - przygody: zmiana wyglądu strony, resety i wyłączanie komputera, ponowne ściąganie artykułów
  - łączenie rozdzielonych wydań



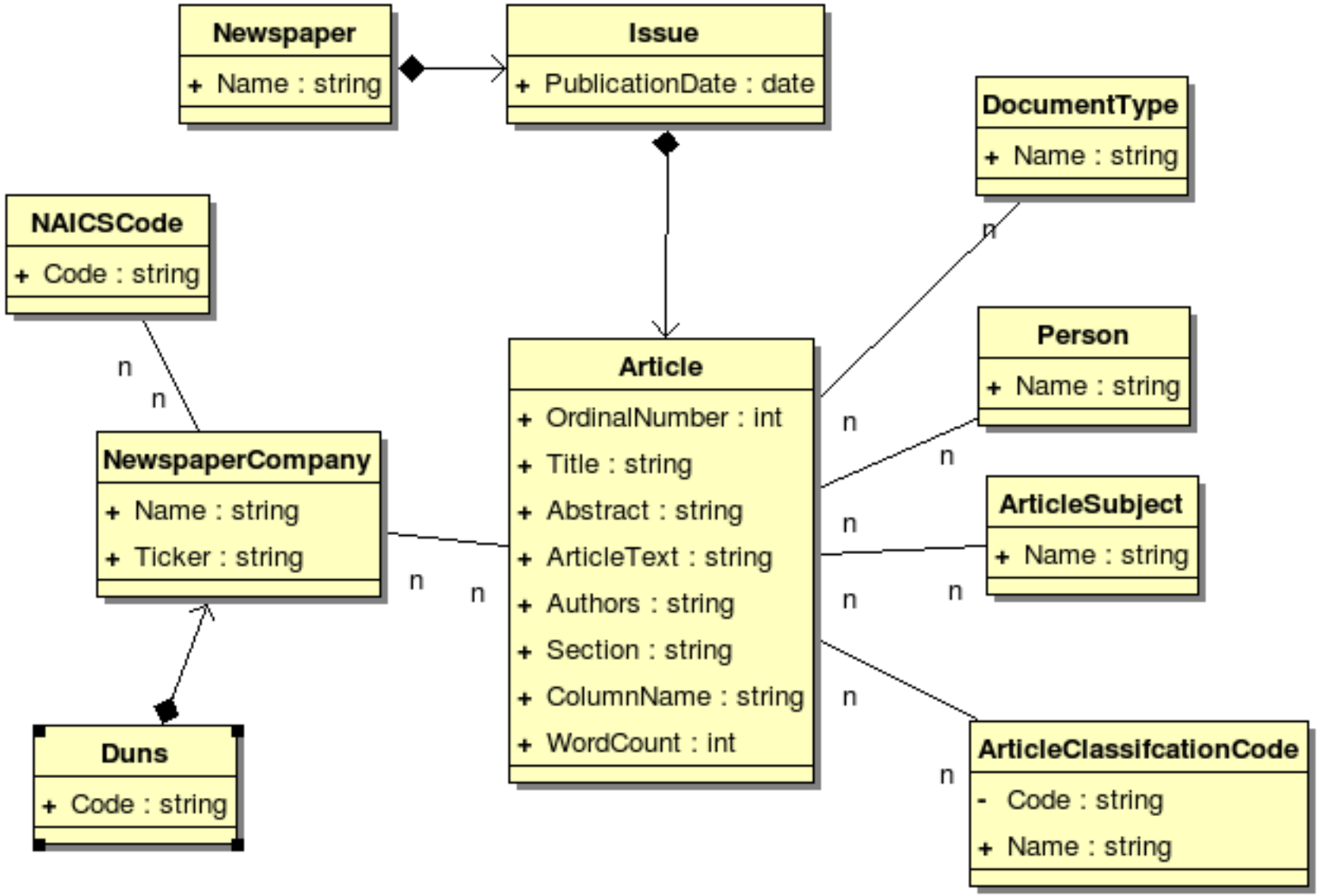
Baza danych

- 2 części
  - notowania akcji, indeksów (z Yahoo.com)
  - artykuły prasowe (Wall Street Journal, Financial Times z ProQuest)

Baza danych – notowania



Baza danych - artykuły



Baza danych podstawowe statystyki

- Artykuły
  - Przedział czasowy: 1997-04-01 – 2007-04-01
  - Rozmiar plików (strony HTML) na dysku: ok. 60GB
- Notowania akcji i indeksów
  - Przedział czasowy: do 2007-06-15
  - Rozmiar plików (strony HTML i CSV) na dysku: ok. 20 MB
- Rozmiar całej bazy danych: ok. 8 GB

Baza danych – notowania podstawowe statystyki

- Indeksów: 73
- Firm: 8672

Baza danych – artykuły podstawowe statystyki

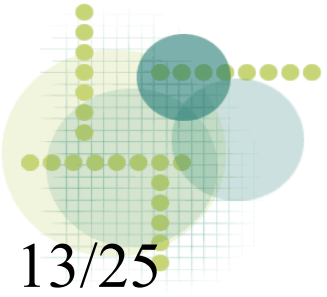
		WSJ	FT
Wydania		2806	3104
Artykuły		422009	765117
Brakujące artykuły		341	9
Artykułów w wydaniu	min	1	1
	max	730	437
	AVG	150,52	246,5
	SD	77,73	52,4
Wybrane puste atrybuty artykułu	tytuł	0	0
	streszczenie	7,23%	13,58%
	tekst	9,55%	~0%
meta-atrybuty		18	14
Dostępność wybranych meta-atrybutów	Subjects	74,19%	62,50%
	Section	25,72%	99,73%
	Companies	48,04%	36,10%
	Document types	73,71%	62,62%
	Column name	21,71%	0,00%
	Classification codes	25,74%	0,00%

Artykuły – czyszczenie danych

- Dane o firmach opisywanych w artykule
  - Łączenie danych o firmie (wg nazwy) dostępnych w różnych artykułach jako meta-atrybuty:
    - Nazwa
    - NAICS (klasyfikacja przemysłowa)
    - Duns (jednoznacznie identyfikuje firmę)
    - Symbol (ticker)
  - redukcja: z 46318 do 42554

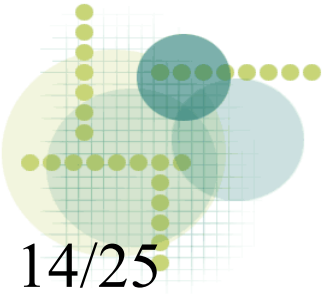
Artykuły – czyszczenie danych

- Artykuł – czyszczenie tekstu
  - usuwanie tabel (i tu np. tekst w tabeli)
  - wstawianie NULL do DB zamiast ""
  - usuwanie wpisu „Copyright ...”
  - zamiana specjalnych znaków HTML-a na zwykłe
  - usuwanie końcówki „; [NUMER]” w tytule



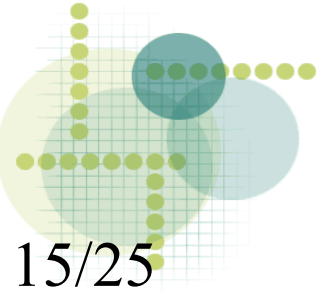
Artykuły – czyszczenie danych

- Spostrzeżenie: niektóre artykuły częściowo powtarzają się w danym wydaniu
  - różnią się meta-atrybutami,
  - jeden jest rozszerzoną wersją drugiego
- Rozwiązanie pierwszego problemu – łączenie (sumowanie meta-atrybutów) tożsamych artykułów



Artykuły – czyszczenie danych

- 2 artykuły są tożsame jeśli poniższy program zwróci „TRUE”:
  - if(text[0] !=NULL AND text[1] != NULL):
    - if(text[0]==text[1]) RETURN TRUE
  - else:
    - if(abstract[0] != NULL AND abstract[1] != NULL):
      - if(abstract[0]==abstract[1]) RETURN TRUE
    - else if(title[0]==title[1]) RETURN TRUE
  - RETURN FALSE
- Redukcja liczby artykułów: z 421863 do 418250

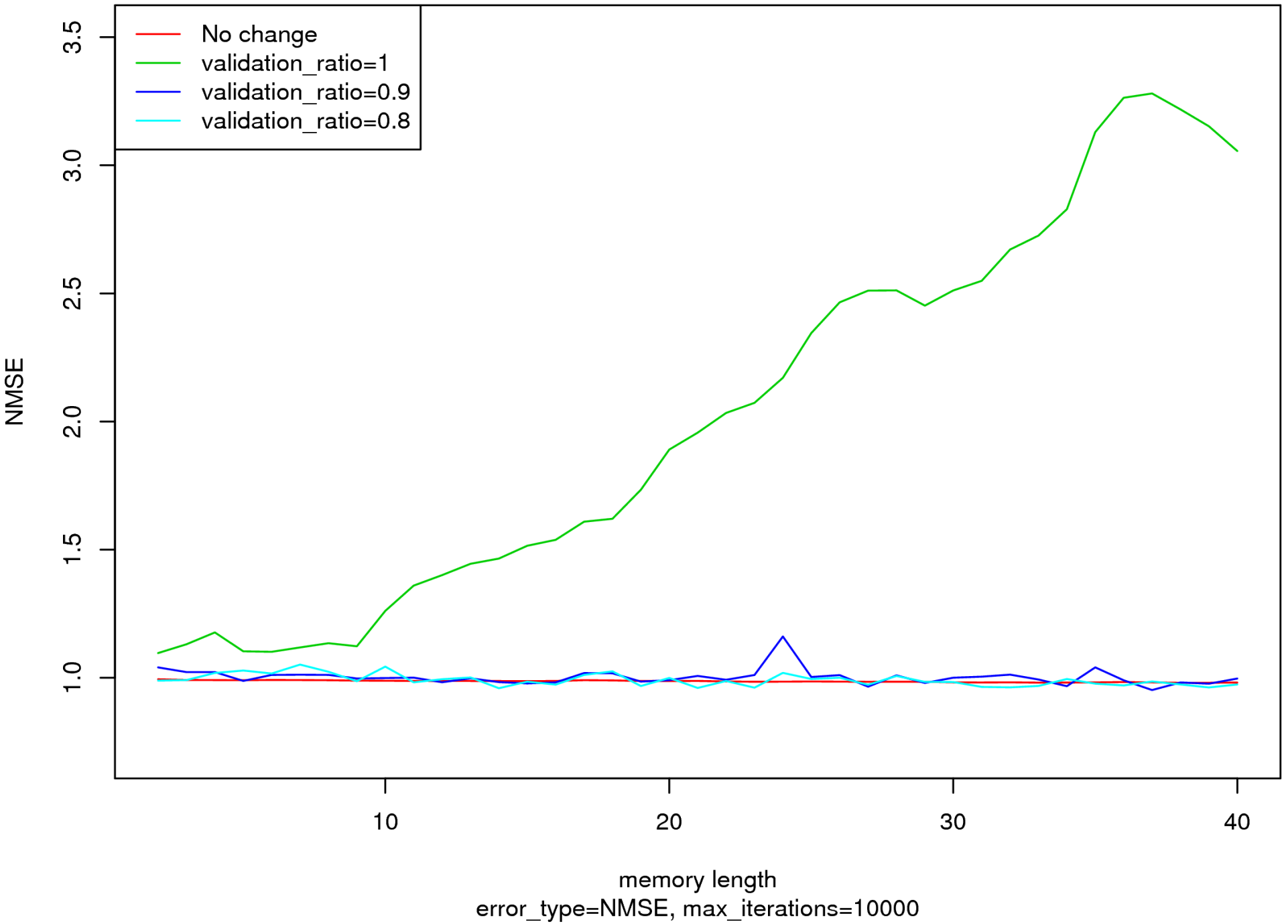


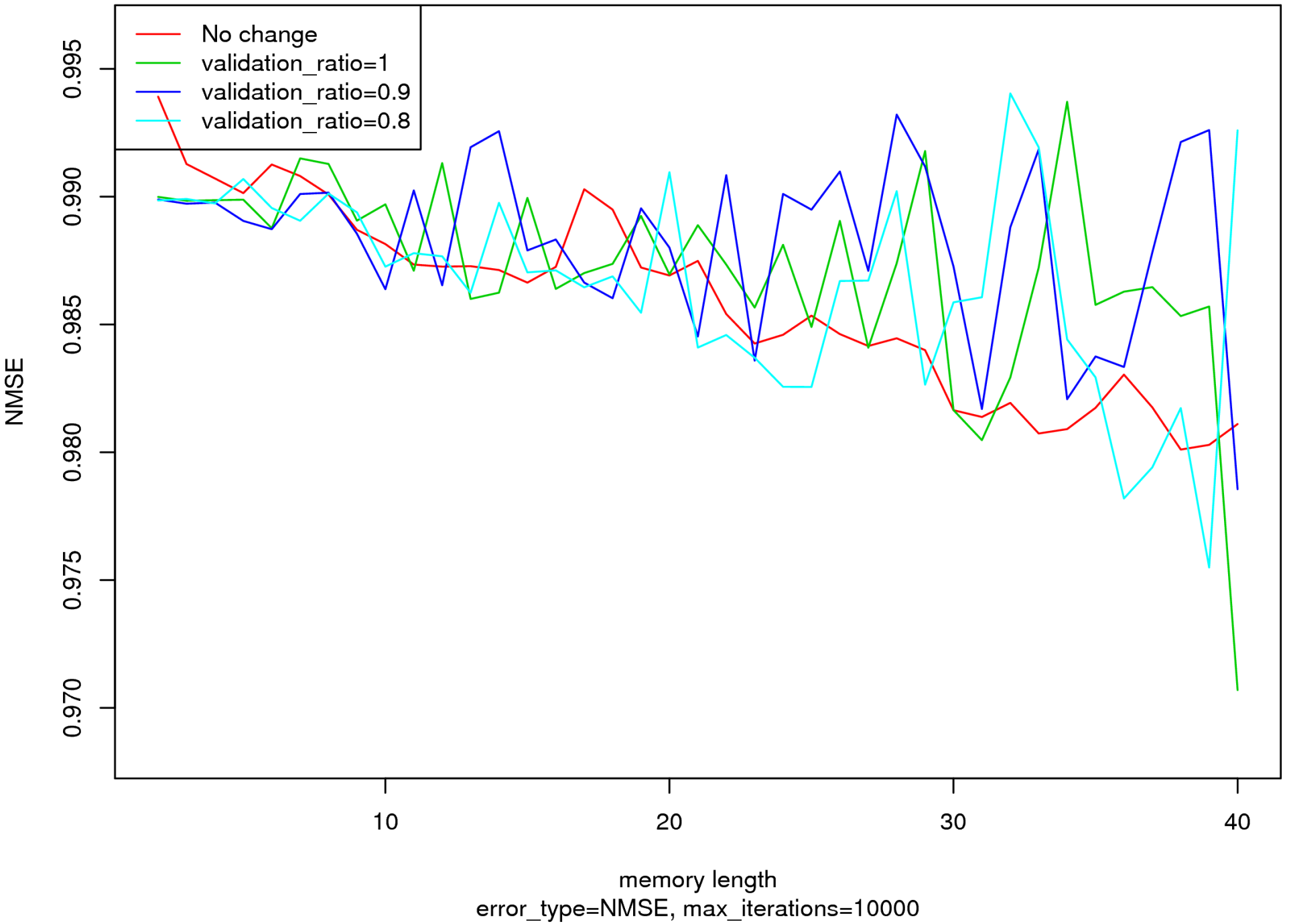
Proste eksperymenty

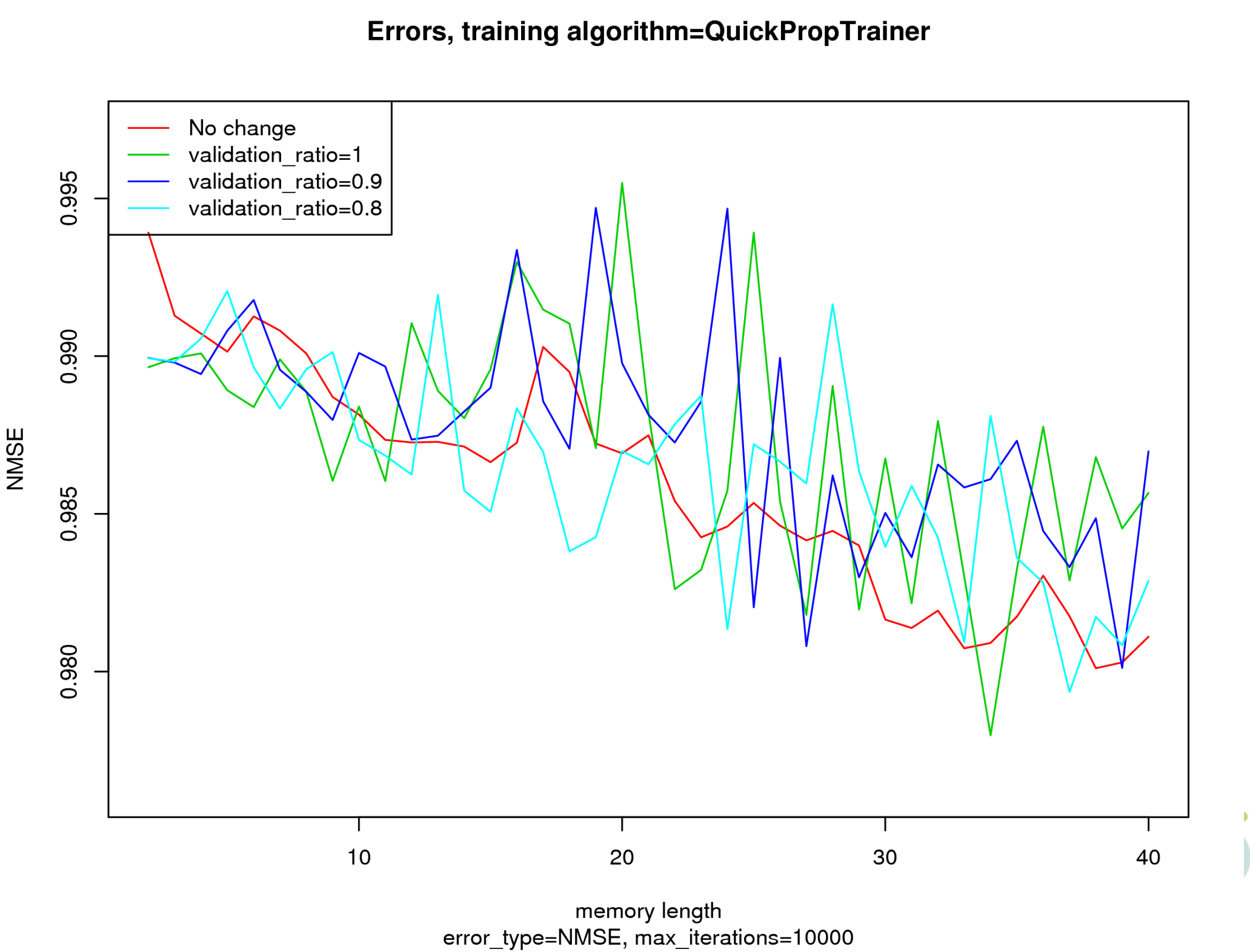
- Dane:
  - Szereg czasowy zwrotów $((v_i - v_{i-1})/v_{i-1})$ ze znormalizowanych kwotowań indeksu S&P500.
  - Przedział czasowy: 2003-04-01 – 2007-04-01
    - od 926 do 1002 elementów szeregu czasowego
  - Przewidywanie zwrotu dzień naprzód
  - zbiory:
    - trenujący: pierwsze 90% szeregu
    - testujący: kolejne 10%

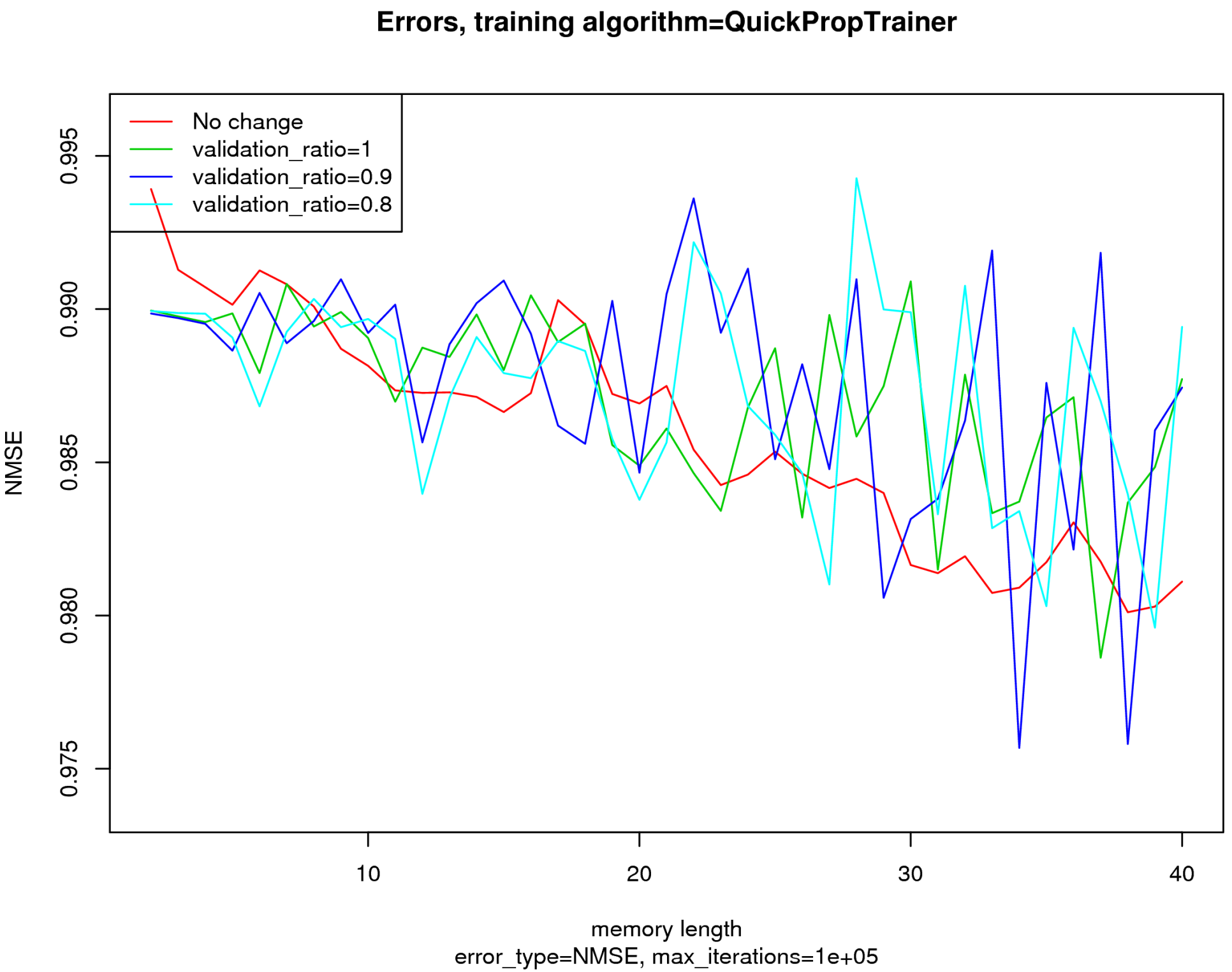
Proste eksperymenty

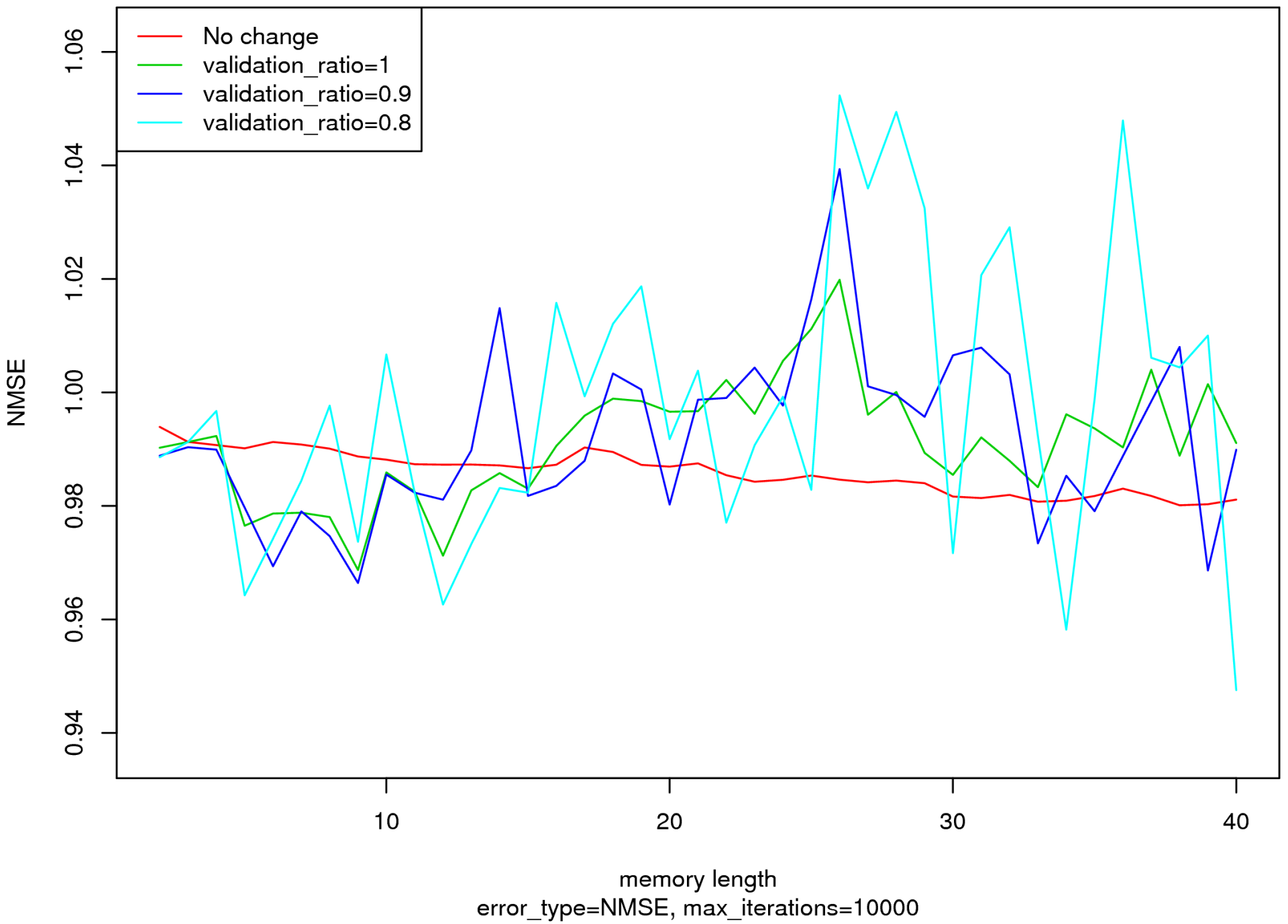
- Parametry:
  - długość „pamięci”
  - wielkość zbioru walidacyjnego
  - maksymalna liczba epok/iteracji uczących
  - algorytm uczący (On-line, batch, Rprop, QuickProp)
- Liczony błąd:
  - NMSE:
    - Y, Y' – wartość oczekiwana i wartość otrzymana z modelu
    - $NMSE(Y', Y) = E((Y' - Y)^2)/E((EY - Y)^2) = MSE/Var(Y)$

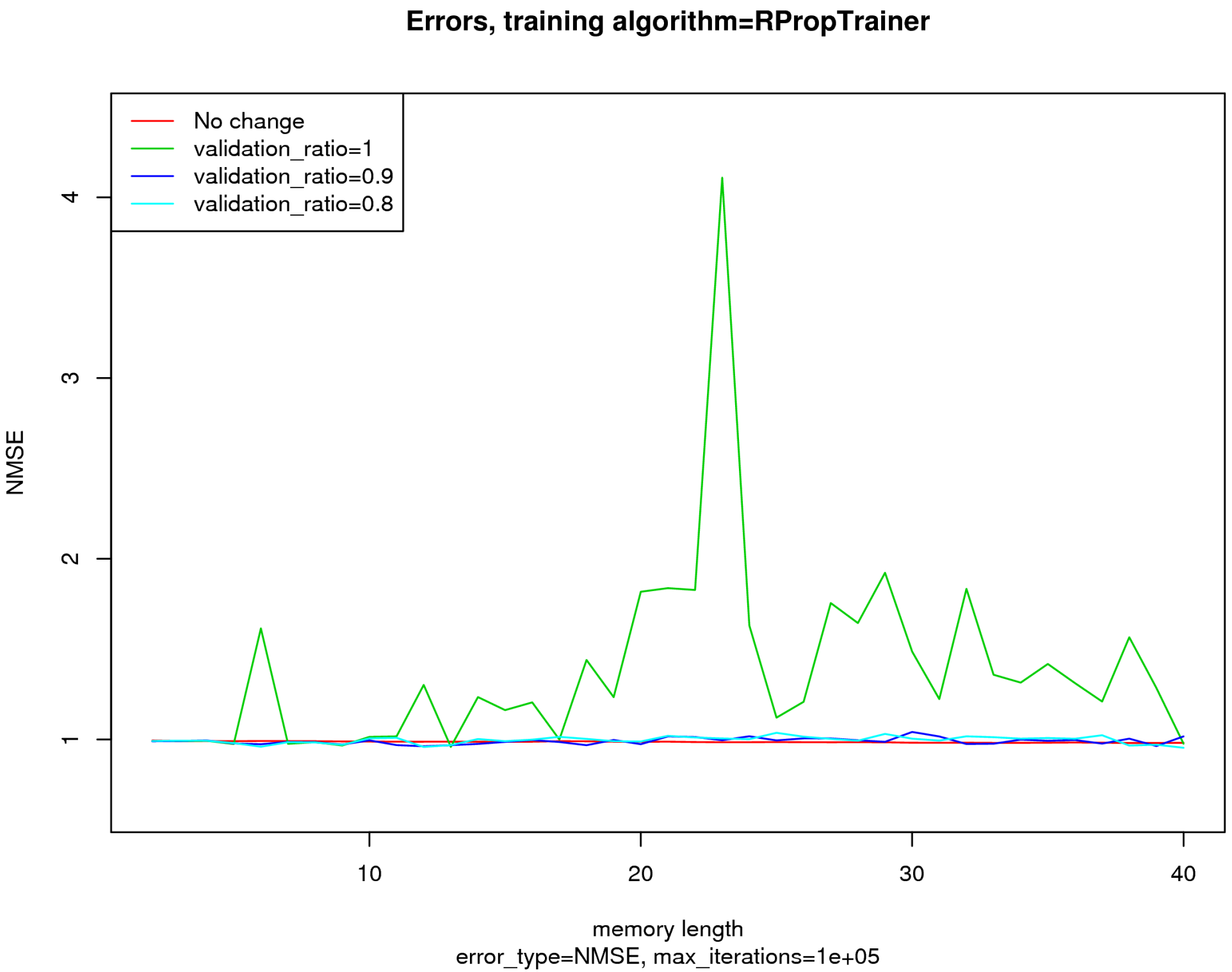

Errors, training algorithm=StandardOnLineTrainer
No change
validation_ratio=1
validation_ratio=0.9
validation_ratio=0.8
NMSE
1.0
1.5
2.0
2.5
3.0
3.5
10
20
30
40
memory length
error_type=NMSE, max_iterations=10000


Errors, training algorithm=StandardBatchTrainer
No change
validation_ratio=1
validation_ratio=0.9
validation_ratio=0.8
NMSE
0.995
0.990
0.985
0.980
0.975
0.970
10
20
30
40
memory length
error_type=NMSE, max_iterations=10000


Errors, training algorithm=QuickPropTrainer
No change
validation_ratio=1
validation_ratio=0.9
validation_ratio=0.8
NMSE
0.995
0.990
0.985
0.980
10
20
30
40
memory length
error_type=NMSE, max_iterations=10000


Errors, training algorithm=QuickPropTrainer
No change
validation_ratio=1
validation_ratio=0.9
validation_ratio=0.8
NMSE
0.995
0.990
0.985
0.980
0.975
10
20
30
40
memory length
error_type=NMSE, max_iterations=1e+05


Errors, training algorithm=RPropTrainer
No change
validation_ratio=1
validation_ratio=0.9
validation_ratio=0.8
NMSE
0.94
0.96
0.98
1.00
1.02
1.04
1.06
10
20
30
40
memory length
error_type=NMSE, max_iterations=10000


Errors, training algorithm=RPropTrainer
No change
validation_ratio=1
validation_ratio=0.9
validation_ratio=0.8
NMSE
1
2
3
4
10
20
30
40
memory length
error_type=NMSE, max_iterations=1e+05

Przyszłość

- Dalsze oczyszczanie danych (może)
  - łączenie b. podobnych artykułów (zmieniona kolejność słów, dodane jedno zdanie itp.)
  - ignorowanie artykułów np. z kolumny humorystycznej gazety
- Pomysły
  - zastosować podstawowe reprezentacje tekstu (TF-IDF)
  - uwzględnić model wyceny akcji (CAPM)
  - użyć predefiniowany słownik ważnych zwrotów
  - przewidywanie cen akcji firm i sektorów przemysłu

Dziękuję za uwagę!