Dmytro Martsynenko 36427

Inteligentna analiza

Analiza dużego zbioru tekstów metodą kNN

W trakcie opracowania projektu utworzono aplikację, która dokonuje analizy zbioru tekstów.

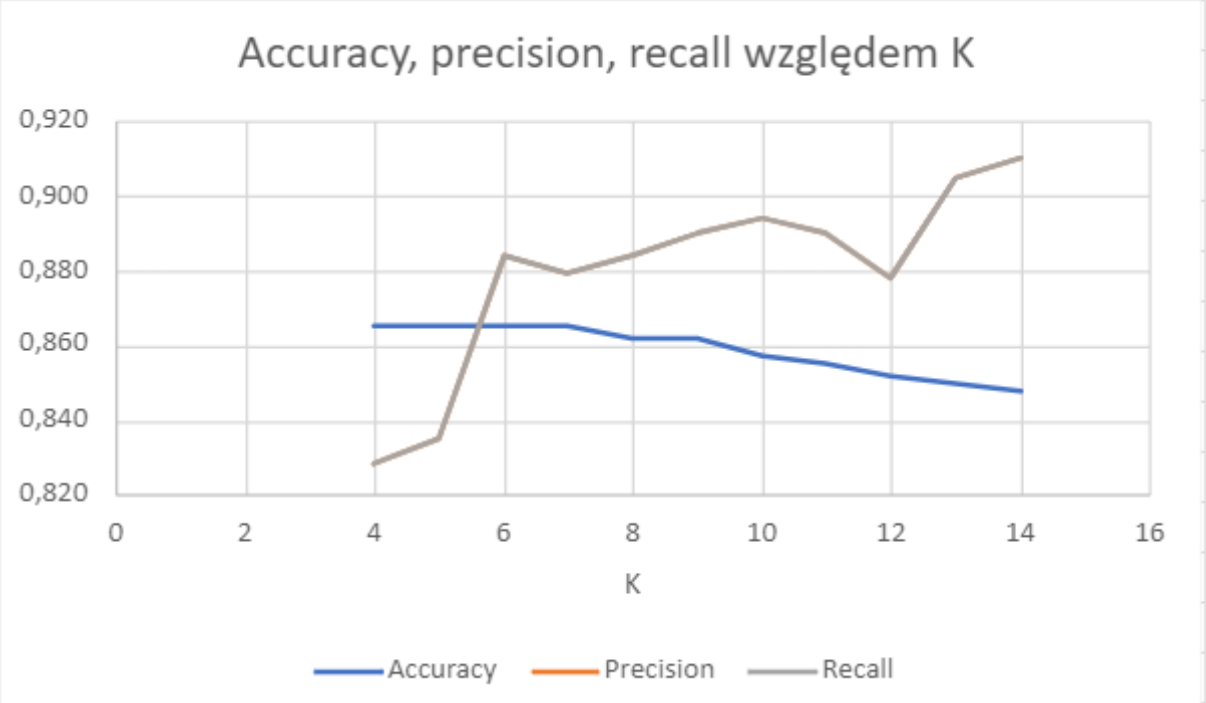
Cele projektu:

1. Porównać wyniki klasyfikacji metody k-NN dla 10 różnych wartości parametru k (wyznaczyć zależność Accuracy od k, przy stałych wartościach innych parametrów).
2. Przy wybranej stałej wartości k wyznaczyć zależność Accuracy od pięciu wartości proporcji podziału zbioru (przy pozostałych parametrach stałych).
3. Wyznaczyć zależność Accuracy od wyboru metryki/miary (przy pozostałych parametrach stałych).
4. Na podstatwie dowolnego wyboru 4-ch podzbiorów cech wskazać, które cechy potencjalnie mają najmniejszy, a które największy wpływ na wyniki klasyfikacji, zwłaszcza na Accuracy (przy innych wartościach stałych).

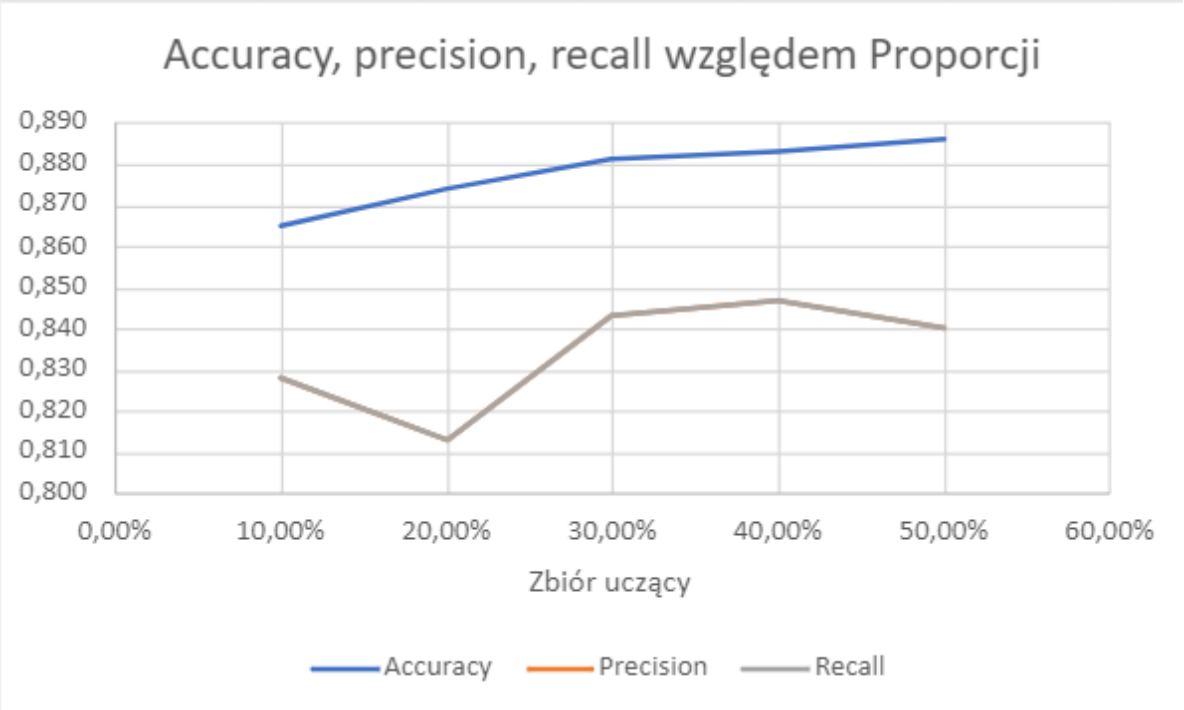
Działanie programu można opisać w następujący sposób:

1. Przeczytaj pliki w postaci XML.
2. Przetwórz dzrzewo XML w obiekt.
3. Wybierz atrybuty Places oraz TEXT.BODY.
4. Utwórz obiekt klasy Article i usuń niepotrzebne znaki w tekscie.
5. Przestemuj teksty, znajdujące się w obiektach tej klasy.
6. Wykorzystaj ScenarioFactory, żeby utworzyć scenariuszy uruchomienia analizy, zgodne z poszczególnymi celami projektu.
7. Utworzone scenariusze wraz z tekstami przekaż do ScenarioRunner
8. Przeprowadź analizę i wyświetl wyniki, ewentualnie zapisz ich do pliku

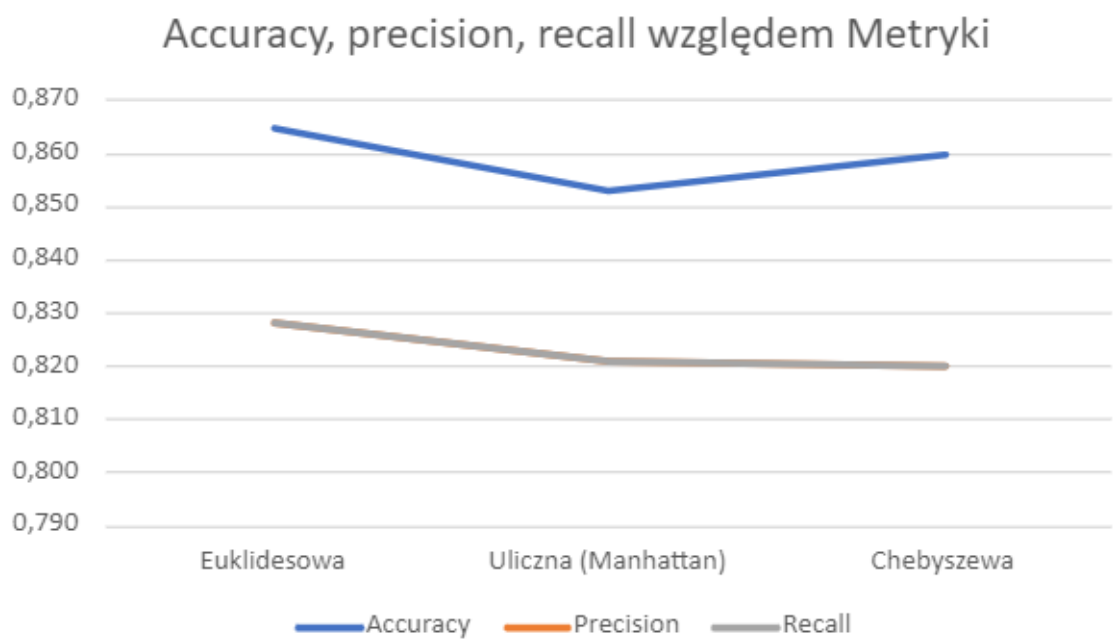
Analiza według wszystkich scenariuszy zajmuje ok. 6 godzin



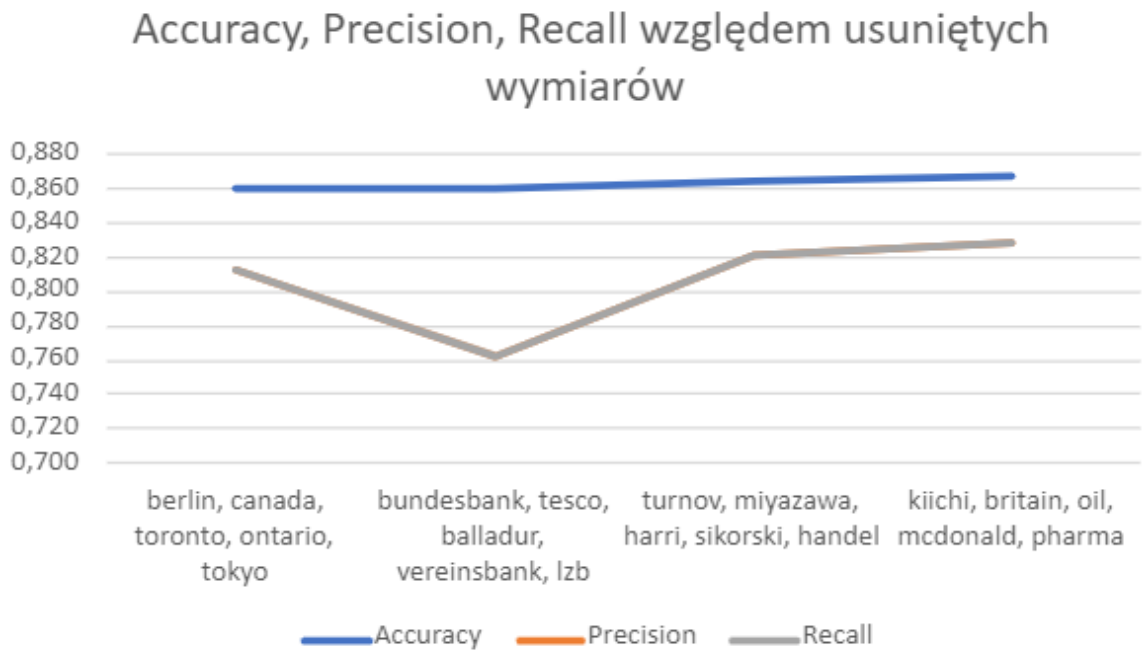
Z powyższego wykresu wynika, że przy zwiększeniu wartości K, Accuracy lekko zmniejsza się, gdy Precision oraz Recall zwiększają się. Są one równe, ponieważ algorytm wyłącznie przypisuje kategorię do tekstu, a nie wybiera kategorię, do której tekst nie należy.



Przy zmianie proporcji zbioru uczącego do zbioru testowego, Accuracy rośnie, Precision wraz z Recall nieco zmniejsza się przy 20% zbioru uczącego, ale po tym rośnie, po 40% ewentualnie maleje.



Wpływ wyboru metryki zmniejsza Accuracy w przypadku metryki Ulicznej, pozostałe prawie nie różnią się.



Usunięcie wymiarów (cech) nie powoduje gwałtownych zmian w analizie, oprócz cech *bundesbank, test, balladur, vereinsbank, lzb,* po usunięciu których zmniejszyły się wartości Precision oraz Recall. Z tego można wywnioskować, że są to cechy które najwięcej wpływają na wyniki analizy, wśród innych wskazanych na wykresie. Te inne są potencjalnie najmniej znaczące.