Rozdział 1

Wstęp

Rozwój internetu przyczynił się do powstania ogromnej ilości danych w postaci tekstowej. W konsekwencji stają się one coraz powszechniej wykorzystywanym źródłem, często cennych, informacji. Dane tekstowe są trudne w analizie z powodu ich objętości, różnorodności oraz podatności na błędy. Dlatego bardzo ważnym elementem ekstrakcji wiedzy z danych tego typu jest ich przetwarzanie.

Jednym z istotnych zagadnień związanych z danymi tekstowymi jest grupowanie. W wielu przypadkach kategoryzowanie dokumentów pod kątem podobieństwa tematycznego znacząco ułatwia wyszukiwanie informacji. Z grupowaniem wiążą się dwa zasadnicze elementy: sposób reprezentacji danych oraz algorytm używany do wykrywania skupień. Oczywistym jest, że gdy chcemy kategoryzować teksty pod kątem tematycznym, to ich reprezentacja powinna zachowywać informacje o temacie. Zazwyczaj przyjmuje się naturalne założenie, że kolejność słów nie ma znaczenia – istotny jest jedynie ich rozkład. Podstawowym podejściem jest charakteryzacja tekstu przy użyciu liczności poszczególnych słów w dokumencie – cały zbiór jest wówczas opisany przez macierz, której ij-ty element to liczność wystąpień j-tego słowa w i-tym tekście. Jednakże taka reprezentacja jest najczęściej nieefektywna - dane opisane w ten sposób mają bardzo duży wymiar (równy liczbie wszystkich unikalnych słów w nich zawartych), a przy tym są one bardzo rzadkie – liczba zerowych elementów macierzy jest bardzo duża (często ponad 90%). Analiza skupień na takim zbiorze danych jest często wysoce nieskuteczna (tzw. klątwa wielowymiarowości), a czasem wręcz niemożliwa z powodu braku zasobów pamięciowych i obliczeniowych. Rodzi się więc potrzeba zmniejszenia wymiaru danych.

Często stosowaną praktyką jest grupowanie słów, by następnie reprezentować teksty przy użyciu wektorów $(n_1, \ldots, n_K)$, gdzie n_i jest licznością słów z i-tej grupy, $i = 1, \ldots, K$, a K liczbą grup słów. Słowa można pogrupować na wiele sposobów. Przede wszystkim można oprzeć się na założeniu, że fleksja danego słowa jest nieistotna, tzn. ważne jest znaczenie wyrazu, a nie jego odmiana, co wydaje się być rozsądnym podejściem w kategoryzacji tematycznej. Stąd niejednokrotnie stosowanym elementem przetwarzania danych tekstowych jest *stemming*, czyli sprowadzenie słowa do jego rdzenia, np. wyrazy `robiący, robiłem, robię` sprowadzają się do słowa `robić`. W najprostszej wersji podejście to wymaga dysponowania słownikiem słów wraz z ich odmianami. Rzadko jednak zawiera on wszystkie istniejące wyrazy, zwłaszcza gdy liczba możliwych odmian jest bardzo duża (jak np. w języku polskim). Ponadto podejście to nie uwzględnia częstych niedoskonałości danych – słownik nie zawiera słów z błędami w pisowni i literówkami, które mogą występować w tekstach. Coraz częściej

przedmiotem analizy są nieformalne teksty lub wiadomości, w których np. piszący świadomie nie używa znaków diakrytycznych. Zachodzi zatem potrzeba radzenia sobie z błędnie zapisanymi wyrazami.

Z wymienionych wyżej powodów często stosowaną praktyką jest grupowanie wyrazów w oparciu o pewne miary podobieństwa określone na przestrzeni napisów (ciągów o dowolnej długości nad pewnym zbiorem skończonym, zwanym alfabetem). Można wyróżnić tutaj dwa podejścia: przyporządkowywanie słów do z góry określonych grup (definiowanych przez formy podstawowe) lub wykrywanie skupień spośród wyrazów występujących w tekstach. Niniejsza praca porusza problem wyboru odległości przy zastosowaniu pierwszego z wymienionych podejść. W literaturze można znaleźć wiele różnych definicji odległości (np. [16, 6, 18]). W głównej mierze opierają się one na porównywaniu liczby wystąpień takich samych sekwencji znaków lub zliczeniu operacji, które przetwarzają jeden napis w drugi. Dodajmy, że miary te znajdują zastosowanie również w innych dziedzinach, m.in. biologii obliczeniowej, przetwarzaniu sygnałów czy korekcie błędnego tekstu.

Celem niniejszej pracy jest zbadanie wpływu doboru odległości na przestrzeni napisów na jakość automatycznej kategoryzacji tematycznej tekstów. Innymi słowy, badana jest jakość grupowania artykułów dla różnych reprezentacji tekstu. Reprezentacje te odnoszą się do różnych grup słów, otrzymanych przy użyciu różnych odległości. Danymi, na których przeprowadzona została analiza jest zbiór artykułów polskiej Wikipedii. W pierwszym etapie grupowane są słowa przy użyciu wybranych odległości. Dalej reprezentujemy artykuły jako liczności występowania poszczególnych grup słów w danym tekście. Na podstawie tak uzyskanych danych przeprowadzamy analizę skupień tekstów. Ocena wyników odbywa się na podstawie otrzymanych grup z prawdziwymi kategoriami przypisanymi do każdego tekstu Wikipedii.

Zauważmy, że użycie polskiej Wikipedii wiąże się to z kilkoma istotnymi wyzwaniami. Po pierwsze język polski jest językiem bardzo złożonym, przede wszystkim ze względu na to, że zawiera dużo odmian (przez przypadki, liczby, osoby, czasy, tryby, strony i inne). Stąd liczba istniejących słów w języku polskim jest bardzo duża. Dalej określenie formy bezokolicznikowej czy mianownikowej często nie jest łatwe, a czasem niemożliwe bez znajomości kontekstu (np. słowo `piła` może być zarówno rzeczownikiem, jak i być odmianą czasownika `pić`). Po drugie zbiór tekstów z polskiej Wikipedii jest względnie duży – ponad milion artykułów, na które składają się ponad dwa miliony unikalnych słów. Przetwarzanie tak obszernego zbioru rodzi potrzebę odpowiedniego zarządzenia danymi (zbudowaniem adekwatnej bazy danych) oraz optymalizacji procesów z powodu ograniczonych zasobów obliczeniowych i pamięciowych. Po trzecie niektóre metody analizy danych nie dają się efektywne stosować na dużych zbiorach danych.

Organizacja niniejszej pracy jest następująca. Rozdział drugi to przegląd odległości na przestrzeni ciągów znaków. Podano ich formalne definicje, jak i przykłady użyć oraz zastosowania. Zostały one podzielone na trzy kategorie: odległości oparte na operacjach edycyjnych, oparte na q-gramach oraz miary heurystyczne. Trzeci rozdział poświęcony jest analizie skupień. Opisane zostały algorytmy k-średnich przy użyciu zarówno metody wsadowej, jak i metod najszybszego spadku. Ponadto w rozdziale tym omówione zostały algorytmy hierarchiczne wraz z różnymi kryteriami odmienności między skupieniami. Co więcej są tam opisane metody oceny jakości podziału na skupienia. Kolejny rozdział dotyczy części praktycznej pracy. Przedstawiony został algorytm zastosowany na analizowanym zbiorze wraz z dokładnym opisem badania. Ponadto w rozdziale czwartym znajdują szczegółowe wyniki. Ostatnia część dotyczy kierunków dalszych prac.

Literatura

[1] Dokumentacja modułu sklearn.metrics języka python. `http://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation`. Dostęp: 2015-12-01.

[2] Wikipedia - wikipedia, wolna encyklopedia. `http://pl.wikipedia.org/wiki/Wikipedia`. Dostęp: 2015-12-01.

[3] Daniel Aloise, Amit Deshpande, Pierre Hansen, Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

[4] Léon Bottou. Stochastic gradient tricks. Grégoire Montavon, Genevieve B. Orr, Klaus-Robert Müller, redaktorzy, *Neural Networks, Tricks of the Trade, Reloaded*, Lecture Notes in Computer Science (LNCS 7700), strony 430–445. Springer, 2012.

[5] Léon Bottou, Yoshua Bengio. Convergence properties of the k-means algorithms. *Advances in Neural Information Processing Systems 7*, strony 585–592. MIT Press, 1995.

[6] Leonid Boytsov. Indexing methods for approximate dictionary searching: Comparative analysis. *J. Exp. Algorithmics*, 16:1.1:1.1–1.1:1.91, 2011.

[7] Fred J. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176, 1964.

[8] N.R. Dixon, T.B. Martin. *Automatic Speech and Speaker Recognition*. IEEE Press book. IEEE Press, 1979.

[9] E. B. Fowlkes, C. L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[10] R. W. Hamming. Error Detecting and Error Correcting Codes. *Bell System Techincal Journal*, 29:147–160, 1950.

[11] Trevor J. Hastie, Robert John Tibshirani, Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2009.

[12] Matthew A. Jaro. *UNIMATCH: A record linkage system: User manual*. United States Bureau of the Census, 1978.

[13] J. Koronacki, J. Ćwik. *Statystyczne systemy uczące się*. Wydawnictwa Naukowo-Techniczne, 2005.

[14] Karen Kukich. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4):377–439, 1992.

[15] G. N. Lance, W. T. Williams. A general theory of classificatory sorting strategies 1. hierarchical systems. *The Computer Journal*, 9(4):373–380, 1967.

[16] V. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17, 1965.

[17] H. Masters. *A study of spelling errors.* Univ. of Iowa Studies in Educ. 4, 1927.

[18] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[19] S. B. Needleman, C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[20] O. Owolabi, D. R. McGregor. Fast approximate string matching. 18(4):387–393, 1988.

[21] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[22] Andrew Rosenberg, Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL)*, strony 410–420, 2007.

[23] Peter Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, 1987.

[24] D. Sankoff, J. B. Kruskal. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.* Addison-Wesley, 1983.

[25] D. Sculley. Web-scale k-means clustering. *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, strony 1177–1178, New York, NY, USA, 2010. ACM.

[26] Peter H Sellers. The theory and computation of evolutionary distances: Pattern recognition. *Journal of Algorithms*, 1(4):359 – 373, 1980.

[27] Esko Ukkonen. Algorithms for approximate string matching. *Inf. Control*, 64(1-3):100–118, 1985.

[28] Esko Ukkonen. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92(1):191 – 211, 1992.

[29] Mark P. J. van der Loo. The stringdist Package for Approximate String Matching. *The R Journal*, 6:111–122, 2014.

[30] T. K. Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics*, 4(1):52–57, 1968. Russian Kibernetika 4(1):81-88 (1968).

[31] Robert A. Wagner, Michael J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.

[32] Robert A. Wagner, Roy Lowrance. An extension of the string-to-string correction problem. *J. ACM*, 22(2):177–183, 1975.

[33] J.H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[34] Anna Wilbik, James M. Keller. A distance metric for a space of linguistic summaries. *Fuzzy Sets and Systems*, 208:79–94, 2012.

[35] William E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research*, strony 354–359, 1990.

[36] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37, 2007.