|  |  |  |
| --- | --- | --- |
| Название | Формула или словесное  описание расчета (по-русски) | Применение (если есть), ссылка на  источник |
| Longest common string (LCS)/  Самая длинная общая строка | , где m и n – длины строк, LCSuff – функция, находящая самый длинный общий суффикс из возможных префиксов S1 и S2 | Алгоритм помогает найти самую длинную подстроку в строке.  Jaccard, P. (1901).  Etude comparative de la distribution florale dans une portion des Alpes et du  Jura. Impr. Corbaz. |
| Damerau -Levenshtein | Это дистанция или строковая мера между двумя строками, которая дает число необходимое для преобразования одной строки в другую. Преобразование происходит с помощью вставки, удаления или замены одного символа, или перестановки двух соседних символов. | Bard, G. V.(2007). Spelling–error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string edit distance metric.  Proceedings of the fifth Australasian symposium on  ACSW frontiers-Volume 68, Australian Computer  Society, Inc., pp. 117–124. |
| Jaro | если m=0,  в других случаях,  где m - число совпадающих символов, t – половина от числа перестановок.  Два символа из s1 и s2 совпадают только если они одинаковы и не далеки от:    dj – мера похожести 2-х строк: если 0 – не похожи, 1 – это одно и тоже. | Jiang, J. J. & Conrath, D. W. (1997). Semantic  similarity based on corpus statistics and lexical  taxonomy. arXiv preprint cmp-lg/9709008. |
| Jaro–Winkler | Расстояние dj между двумя строками s1 и s2:    где p – шкала префиксов, предоставляющая рейтинг строк, совпадающих по началу длины префикса l, dj – расстояние Jaro | Этот метод подходит для коротких строк, например, имен.    Формула: Krause, E. F. (1973).  Taxicab geometry. The Mathematics Teacher, Vol. 66, No. 8, pp. 695–706.  Статья:Winkler, W. E. (2006). Overview of record linkage and current research directions. Bureau of the Census. |
| Needleman-Wunsch | В статье нет | Динамический алгоритм, используемый в биоинформатике для выравнивания белковых последовательностей. Это также используется как метод оптимального соответствия и общая техника выравнивания.  Needleman, S. B. & Wunsch, C. D. (1970).  A general method applicable to the search for  similarities in the amino acid sequence of two  proteins. Journal of molecular biology, Vol. 48, No. 3, pp. 443–453. |
| Smith-Waterman | Чтобы измерить сходство, его сравнивают внутри строк, а затем оптимизируют сходство | Не используется для больших проблем, потому что имеет кубическую сложность.  Smith, T. F. & Waterman, M. S. (1981).  Identification of common molecular subsequences.  Journal of molecular biology, Vol. 147, No. 1 |
| n–gram | Вероятностная модель, используемая для предсказывания последующего члена последовательности из (n-1) символов или слов. | Используется в вычислительной биологии (анализ биологических последовательностей ), в сжатии данных, компьютерной лингвистике и в вычислительных теориях.  Особенность – простота и расширяемость.  Barr ´ Labaka, G. (2010). Plagiarism detection across  distant language pairs. Proceedings of the 23rd International Conference on Computational Lin-  guistics, Association for Computational Linguistics,  pp. 37–45. |
| syntactic n–gram | Элементы берутся не в том порядке, в котором они появляются в тексте, а в том, в котором они появляются в синтаксическом дереве. | Применяются в во всех задачах, в которых можно использовать n-gram  1) Sidorov,G.(2013). Non-linear construction of n-grmas in computational lingusitics: Syntactic,  filtered, and generalized n-grams. SMIA, Mexico.  2) Sidorov, G. (2014). Should syntactic n-grams contain names of syntactic relations? International  Journal of Computational Linguistics and Applica-  tions, Vol. 5, pp. 139–158.  3) 41 Posadas-Dur´, J.Markov  ´H., Sidorov, G., Batyrshin, I., Gelbukh, A., &  Pichardo-Lagunas, O. (2015). Syntactic n-grams  as features for the author profiling task. Working  Notes Papers of the CLEF 2015 Evaluation Labs,  volume 1391 of CEUR Workshop Proceedings,  CEUR. |

2.