ИТвПС Задание 19 октября.

Исмакова, Долгушин ИСТ189-1

3.3.1. Измерение сходства на основе символов

|  |  |  |
| --- | --- | --- |
| 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. |
| Damerau -Levenshtein/  Дамерау - Левенштейн | 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 / Жаро | 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. |  |
| Jaro–Winkler / Жаро - Винклер | 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.    Winkler, W. E. (2006). Overview of record linkage and current research directions. Bureau of the Census. |
| Needleman-Wunsch /  Нидлеман - Вунш | 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 /  Смит - Ватерман | 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 /  синтаксический 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. |

3.3.2. Измерение сходства на основе Термов

|  |  |  |
| --- | --- | --- |
| Название | Формула или словесное  описание расчета (по-русски) | Применение (если есть), ссылка на  источник |
| Block Distance  (др. названия: city block distance, Snake distance, Manhattan distance, Manhattan length, L1 distance, Taxicab distance)/  Расстояние между блоками | Дистанция block, d1, двух векторов p и q это | Используется, чтобы найти расстояние между двумя точками.  Krause, E. F. (1973). Taxicab geometry. The Mathematics Teacher, Vol. 66, No. 8, pp. 695–706 |
| Cosine Similarity/  Косинусное сходство | Косинусное сходство двух ненулевых векторов могут быть измерены через Евклидово скалярное произведение.    Косинусное сходство cos(Θ) двух векторов A и B  , где Ai и Bi это составляющие A и B | Используется в интеллектуальном анализе (data mining) данных для поиска сплоченности  между ними (cohesion).  Tan, P.-N. et al. (2006). Introduction to data mining. Pearson Education India. |
| Soft Cosine Similarity/  Мягкое косинусное сходство | , где sij это значение из матрицы сходства между признаками (features) I и j. | Sidorov, G., Gelbukh, A. F., Gomez-Adorno, H., ´ & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. Computacion y Sistemas ´ , Vol. 18, No. 3, pp. 491–504. |
| Sorensen–Dice index  (Др. названия: Sorensen index, Dice’s coefficient)/  индекс Соренсена - Дайса | Было принято, чтобы находить присутствие или отсутствие данных двух множеств    Где |X| и |Y|, это количество элементов в двух множеств, и QS – фактор подобия, колеблющийся от 0 до 1.  Когда используется для нахождения сходства строк S1 и S2, то коэффициент может быть рассчитан, как биграммы, как представлено:    , где nt – число биграм строк и ns1, ns2 – число биграм в строке S1 и S2. | Используется, чтобы найти сходство двух примеров.  Dice, L. R. (1945). Measures of the amount of ecologic association between species. Ecology, Vol. 26, No. 3, pp. 297–302.  Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. Biol. Skr., Vol. 5, pp. 1–34. |
| Euclidean Distance/  Расстояние Евклида | Расстояние Евклида между двумя точками s and t это | Это дистанция между двумя точками в Евклидовом пространстве. |
| Jaccard Index  (Др. названия: Jaccard similarity coefficinet)/  Индекс Жаккара | Он определяется следующим образом | Эта статистика используется, чтобы найти сходство и разнообразие (diversity) между двумя конечными множествами.  Еще используется для нахождения несходства (dissimilarity) между двумя конечными множествами, и получается вычитанием индекса Жаккара из единицы, обозначается как dj.  Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Impr. Corbaz. |
| Simple Matching Coefficient (SMC)/  Коэффициент простого соответствия | Он принимает объекты как набор из n двоичных атрибутов и SMC от А и B -  ,  Где a00 представляет все атрибуты как 0 в А и B; и a10 представляет все атрибуты как 1-цы и 0-ли в А и B; и a10 представляет все атрибуты как 0-ли и 1-цы в А и B; и a11 представляет все атрибуты как 1-цы в А и B. | Данная статистика используется для нахождения сходства и разнообразия (diversity) между двумя объектами |
| Overlap Coefficient  (др. название Szymkiewicz-Simpson coefficient)/  Коэффициент пересечения | Пересечение между двумя множествами определяется как    Коэффициент пересечения равен 1, когда множество A это подмножество множества B. | Является также мерой сходства, связанной с индексом Жаккара. |

Задание 2

Раздел 7: применения семантического сходства текстов.

В каких областях используются меры сходства текстов?

Ответ:

Меры сходства текстов используются в **Биомедицинской информатике,** где для разработки биомедицинской онтологии, а именно Генной онтологии, используется семантическое сходство**, Геоинформатике,** где мера подобия используется для поиска сходств между онтологиями типов географических объектов, и в **Обработке Естественного Языка (NLP),** где меры сходства могут играть важную роль прямо или косвенно в таких полях как: сентиментный анализ, понимание естественного языка и машинный перевод.