Приведен пример, что у нас есть два текста: (1) play, game, (2) player, gamer. Он определяет 4-мерный VSM со следующими функциями: play, player, game, gamer. У нас есть два вектора a и b: a = [1,0,1,0] и b = [0,1,0,1]. Традиционное cosine similarity этих двух векторов равно 0. Но если принять во внимание сходство слов, то оказалось, что эти векторы весьма схожи. cosine similarity не занимается изучением сходства между признаками, изначально считается что все признаки разные. Если же изменить способ вычисления подобия в модели векторного пространства с учетом подобия признаков, то мы получим soft cosine measure (в статье рассматривается модификация с помощью расстояния Левенштейна). То есть в soft cosine measure берут и сравнивают новые признаки полученные, скажем, взяв среднее значение двух признаков одного и того же вектора, умноженное на сходство этих двух признаков. Расстояние Левенштейна - это количество операций (вставок,удалений, перестановок), необходимых для преобразования одной строки в другую. В данном случае расстояние Левенштейна является хорошей мерой для сравнения строк, но можно использовать и другие меры. Итак, если нашими объектами являются тексты, то традиционными признаками являются слова, n-граммы или синтаксические n-граммы, и их соответствующие значения основаны на мере tf-idf. В случае расстояния Левенштейна, если мы используем n-граммы или синтаксические n-граммы, то есть две возможности для сравнения строк: непосредственно сравнивать символьные преобразования или рассматривать каждый элемент n-грамов как единицу сравнения. 8 проведенных экспериментов. Значения получены с использованием традиционной косинусной меры и двух мягких (soft cosine measure) косинусных мер с обоими вариантами расстояния Левенштейна: измеренными в символах и в элементах n-граммов. Это показывает, что soft cosine measure дает лучшие результаты в большинстве экспериментов. Были только две системы, которые не могли достичь лучшей производительности, чем traditional cosine. Сама суть понятна, а вот привести разобрать свой пример самостоятельно затруднительно..