**Косинусное сходство**

Для того чтобы понять что это такое начнем с малого.

Документ в векторной модели рассматривается как неупорядоченное множество термов. Термами в информационном поиске называют слова, из которых состоит текст, а также такие элементы текста, как, например, *2010*, *II-5* или *Тянь-Шань*.

Ну что же такое векторная модель?

**Векторная модель** — представление коллекции документов векторами из одного общего для всей коллекции векторного пространства.

Векторная модель является основой для решения многих задач информационного поиска, как то: поиск документа по запросу, классификация документов (в данному случае это что делается в задачи про сходство строк, только у нас в роде документов выступают предложения из текста), кластеризация документов.

Различными способами можно определить вес терма в документе — «важность» слова для идентификации данного текста.

*Например:*

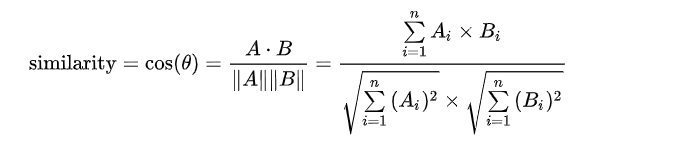
Можно просто подсчитать количество употреблений терма в документе, так называемую частоту терма, — чем чаще слово встречается в документе, тем больший у него будет вес. Если терм не встречается в документе, то его вес в этом документе равен нулю.

Если теперь для некоторого документа выписать по порядку веса́ всех термов, включая те, которых нет в этом документе, получится вектор, который и будет представлением данного документа в векторном пространстве. Размерность этого вектора, как и размерность пространства, равна количеству различных термов во всей коллекции, и является одинаковой для всех документов.

Представление в виде вектора выглядит вот так:

*dj = (a1j , a2j, … , anj )*

*dj* - векторное представление *j*-го документа  
*aij* -вес *i*-го терма в *j*-м документе  
*n* - общее количество различных термов во всех документах коллекции  
  
Вообще если мы располагаем таким представлением для всех документов, можно, например, находить расстояние между точками пространства и тем самым решать задачу подобия документов — чем ближе расположены точки, тем больше похожи соответствующие документы.  
  
С основными терминами познакомились теперь перейдем к главному.  
  
**Косинусное сходство** — это мера сходства между двумя векторами векторного пространства, которая используется для измерения косинуса угла между ними.



Вот такая формул у косинусного сходства.

Как уже говорилось выше с помощью этой формулы сможем определить близкие по смыслу слова в среди большого количества предложений. (наша задача)  
  
**Аппроксимация функции  
  
Аппроксимацией** или по другому приближением   
**функции** f(x) называется нахождение такой **функции** (аппроксимирующей **функции**) g(x), которая была бы близка заданной. Критерии близости **функций** могут быть различные. В случае если приближение строится на дискретном наборе точек, аппроксимацию называют **точечной или дискретной.**

Аппроксимация позволяет исследовать числовые характеристики и качественные свойства объекта, сводя задачу к изучению более простых или более удобных объектов (например, таких, характеристики которых легко вычисляются или свойства которых уже известны).  
  
**Остаточный член** — разность между заданной функцией и функцией её аппроксимирующей. Тем самым оценка остаточного члена является оценкой точности рассматриваемой аппроксимации. Этот термин применяется, например, в формуле ряда Тейлора.  
  
Вообще рядом с термином аппроксимации функции стоит термин интерполяции функции и не спроста.

Наиболее часто встречающим видом точечной аппроксимации является **интерполяция** – нахождение промежуточных значений величины по имеющемуся дискретному набору известных значений.  
  
*2 определение*

**Интерполяция** - это метод нахождения неизвестных промежуточных значений некоторой функции по имеющемуся дискретному набору ее известных значений. Типичным примером такой функции является временной ряд, значения которого – это наблюдения, зафиксированные через определенный интервал времени.**Интерполяционным многочленом** называют алгебраический многочлен степени n - 1, совпадающий с аппроксимируемой функцией в n выбранных точках. Погрешность аппроксимации функции f(x) интерполяционным многочленом степени n-1, построенным по n точкам, можно оценить, если известна ее производная порядка n.