Сравнение ответов в косинусной мере

Алгоритм

1. Удаляем пустые ячейки и строки из ответов без нумерации на риск, описание и мера.

2. Используем **меру TF-IDF** для представления ответов участников и «идеального» ответа в виде числовых векторов, отражающих важность использования каждого слова из некоторого набора слов (количество слов набора определяет размерность вектора) в каждом документе. Подобная модель называется векторной моделью и даёт возможность сравнивать тексты, сравнивая представляющие их вектора в какой-либо метрике (евклидово расстояние, косинусная мера, манхэттенское расстояние, расстояние Чебышёва и др.).

Пример: Если документ содержит 100 слов, и слово «заяц» встречается в нём 3 раза, то частота слова (TF) для слова «заяц» в документе будет 0,03 (3/100). Вычислим IDF как десятичный логарифм отношения количества всех документов к количеству документов содержащих слово «заяц». Таким образом, если «заяц» содержится в 1000 документах из 10 000 000 документов, то IDF будет равной: log(10 000 000/1000) = 4. Для расчета окончательного значения веса слова необходимо TF умножить на IDF. В данном примере, TF-IDF вес для слова «заяц» в выбранном документе будет равен: 0,03 × 4 = 0,12.

3. Исходя из 2-го пункта сравниваем ответы с «идеальным» ответом в **косинусной мере**. Полученный результат от 0 до 1 записываем в файл.

В случае информационного поиска, косинусное сходство двух документов изменяется в диапазоне от 0 до 1, поскольку частота терма (веса TF-IDF) не может быть отрицательной. Угол между двумя векторами частоты терма не может быть больше, чем 90°.