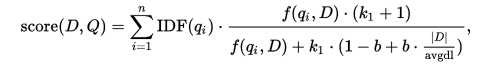
**Text summarization.**

В рамках данной задачи необходимо было представить программу, которая выполняет краткое изложение переданного на вход текста.

Для достижения этой цели была использована python библиотека gensim (<https://radimrehurek.com/gensim/tutorial.html>). Gensim использует алгоритм TextRank с метрикой BM25.

TextRank каждому предложению присваивает метрику называемую прочностью соединения, которая ставится в соответствие количеству слов в предложении. Прочность соединения вычисляется по BM25 алгоритму и находится по следующей формуле.



Где **Q —**  предложение, состоящее из слов q\_1..q\_n.

**D —** документ (в нашем случае текст статьи).

**f(q\_i, D) —** частота слова q\_i в документе

**|D| -** длина документа (количество слов в нем).

***Avgdl —*** средняя длина документа.

**k\_1** и **b** — свободные коэффициенты (приблизительно равны 2.0 и 0.75 соотвественно).

**IDF(q\_i) —** обратная документная частота для слова q\_i.

Таким образом, имеем граф, узлы которого соответствуют прочности соединения для каждого предложения и чем выше данная оценка, тем более ценным считается предложение.

В функции gensim.summarize() присутствует параметр *ratio,* который устанавливает степень сжатия статьи. Чем выше данные параметр, тем более подробное изложение получается и тем больше времени необходимо для выполнения.

Также в программе присутствует функция *keywords* для подсчета и вывода наиболее популярных слов в файле.

На вход программе передается документ, каждая строчка которого содержит тело статьи, для которой необходимо сделать саммари. На выходе получается два файла: *summary* — каждая строчка которого содержит краткое изложение статьи и *keywords* - каждая строчка которого содержит наиболее значимые слова в документе.

Для оценки качества саммаризации используется метрика **ROUGE-n** (*Recall-Oriented Understudy for Gisting Evaluation*,<http://anthology.aclweb.org/W/W04/W04-1013.pdf>)**,** где n- это количество последовательных слов, используемых для оценки. Данная возможность не была имплементирована в данной программе и планируется к реализации в дальнейшей разработке.