# Извлечение ключевых слов

В качестве стоп-слов использовались все найденные в топик-стартере предлоги, союзы, междометия, глаголы, частицы и вводные слова. Выделение стоп-слов производилось экспертным способом. Каждый алгоритм предлагал список пар «ключевое слово – релевантность», отсортированный по убыванию значения релевантности. Для каждого алгоритма в качестве ключевых слов для дальнейшей обработки выбирались слова с релевантностью не меньшей 90% квантили всего списка.

Для извлечения ключевых слов из топик стартера было выбрано 3 алгоритма:

1. Модифицированный RAKE. Представляет собой скрипт на Python. По умолчанию осуществляет выборку наиболее значимых фраз (фраза определяется как набор слов, разделенных стоп-словами). Для каждого слова во фразе вычисляется его частота встречаемости, а также дополнительная характеристика, которая учитывает количество слов во фразе. На последнем шаге рассчитывается значимость каждой фразы на основе рассчитанной ранее дополнительной характеристики. Для скоринга ключевых слов была удалена последнюю часть, которая осуществляет скоринг фраз, тем самым был оставлен непосредственно скоринг ключевых слов. Для ликвидации проблемы с окончаниями и суффиксами в русских словах, была осуществлена обрезка всех рассматриваемых слов в зависимости от их длины.
2. Модифицированный RAKE с учетом экспериментальных функций. Полностью повторяет пункт 1, за исключением того, что были раскомментированы строки, помеченные как экспериментальные функции. Как выяснилось позднее, это никак не повлияло на качество выделения ключевых слов по сравнению с п.1, что лишний раз было подтверждено графиками NCG.
3. Скоринг ключевых слов на основании их частоты употребления. Полностью повторяет п.1, за исключением того, что алгоритм останавливается на вычислении частоты встречаемости каждого укороченного слова. Релевантностью считается то, сколько раз встречается слово в топикстартере.

На основе частоты встречающихся слов - из того же самого RAKE скрипта было удалено все после расчета частоты встречающихся слов. Именно этот алгоритм выдал, на наш взгляд, наиболее релевантные к теме всего форума, ключевые слова.