1. Текстовая релевантность (maxfreq – частота самого частого слова, которая имеет смысл длины документа).

2. Priority bonus, приоритет 7 - текстовый приоритет. Фактор бинарный, имеет значение 0 для всех однословных запросов, и значение 1 практически для всех двух и более словных, кроме очень маленького количества ответов, для которых нет ни одной ссылки, прошедшей кворум, и текст тоже не прошел кворум.

3. Приоритет strict для TR - текстовый приоритет - есть все слова запроса где-то в документе (при этом они проходят контекстные ограничения запроса, например, оба слова д.б. в одном предложении).

4. Приоритет phrase для TR - текстовый приоритет - есть все слова запроса подряд в документе.

5. Наличие точной фразы (текста запроса) в заголовке (если точнее, в первом предложении документа). Контекстные ограничения и стоп слова учитываются в точности как в TRp2, т.е. factor[8] minors factor[5]

6. Встретился участок, прошедший кворум, в котором все словопозиции обозначены как имеющие релевантность BEST_RELEV (заголовок или meta keywords).

7. Длинный документ (чем длиннее документ, тем больше значение фактора).

8. Hitweigt - вариант текстовой релевантности, в которой веса всех хитов считаются равными (т.е. не учитывают надбавки за title и за близость слов). При этом соответствующие хиты должны пройти ограничения синтаксического колдунщика, т.е. можно считать, что фактор TRhitw равен 0 тогда и только тогда, когда SoftAndOk равен 0

9. Длинный текст без ссылок.

10. Фактор про число refines. В языке запросов есть фича user refines ('слово, перед которым стоит знак процентика'). По задумке это означает что-то вроде 'хорошо бы, чтобы слово в документе было'. Единственное известное ((http://staff.yandex-team.ru/gulin Андрей Гулин)) ценное использование данной фичи - это запрос [%официальный %сайт НазваниеФирмы]. Пользователям данная фича неизвестна, т.к. не описана ни в какой документации. Планируется, что она исчезнет из языка запросов, но в колдунщике слова с приоритетом USER_REFINE останутся. Фактор говорит о том, сколько максимум USER_REFINE-слов одновременно встречалось в рамках единого попадания в кворум. Считается, что их от 0 до 3 (если >3, то считается, что 3). Это число мапится в полуинтервал [0,1)

11. Число, на которое умножаются некоторые линковые факторы (именно, факторы номер 6, 7, 47, 66), если текстовая релевантность 0, и ссылок мало

12. В текстовой релевантности произошло совпадение леммы.

13. Dssm модель, обучена на переформулировки, в документной части использует релевантные запросу предложения
14. TR деленный на куб количества слов в запросе и преобразованный стандартным remapTR.
15. Язык документа - русский.
16. Простой BM25 по тексту.
17. Простой BM25 по парам слов - берем все пары слов запроса и считаем число их вхождений в текст документа. В качества веса пары используем сумму весов слов. Комм Не работает, если в запросе есть стоп-слово
18. BM25 от количества предложений в документе, в которых встречается.
19. BM25 по словам только в заголовке.
20. BM25 по словам только с high rel-битиками ('значимым', с выделением (<b> итп)).
21. У документа нет TR.
22. наличие пар слов по точной форме
23. количество предложений, в которых встречается много слов по точной форме
24. наличие слов в заголовке по точной форме
25. BM25 по точной форме
26. Простой BM25 по точной форме.
27. наличие пар слов с учетом синонимов (>=TxtPair)
28. количество предложений, в которых встречается много слов с учетом синонимов
29. наличие слов в заголовке с учетом синонимов
30. BM25 с учетом синонимов
31. Простой BM25 с учетом синонимов.
32. Относительная частота слов запроса в ссылках (1 - слова запроса часто встречаются в ссылках, 0.3 - редко); если точнее, значение этого фактора пессимизируется при условии: TR=0 && LR=0 && (нет ни одной ссылки со всеми словами запроса) && (не прошёл кворум) && (в тексте встречается хотя бы одна пара слов запроса)

33. Документ прошел softand по ограничениям синтаксического колдунщика. Только для документов, имеющих текстовую релевантность. Для однословных запросов всегда 1.
34. Качество текста. Считается по довольно сложной формуле
35. Качество текста (классификатор Алексеева)
36. Длина документа в предложениях
37. Тип документа - HTML
38. документ из порно-кишки
39. фэйковый документ
40. коммерческая страница (классификатор Савина)
41. в документе нет всех слов запроса (с точностью до синонима)
42. процент слов запроса в документе (с точностью до синонима)
43. в документе есть все слова запроса (с точностью до синонима)
44. TR по парам слов запроса в обратном порядке
45. TR по парам слов запроса через одно слово в текстах
46. процент всех слова запроса в тексте (с точностью до формы)
47. в документе есть все слова запроса (с точностью до формы)
48. Длина текста страницы в словах TLen = Map(число слов, 1/400), где Map(x, y) = x*y / (1 + x*y)
49. Длина максимального совпадения форм в тексте и запросе
50. Вес максимального совпадения форм в тексте и запросе
51. Длина максимального совпадения по лемме в тексте и запросе
52. Вес максимального совпадения по лемме в тексте и запросе
53. Варианты соответствующих факторов с учетом стоп слов
54. TR лучшего пассажа - насколько качественный сниппет может получиться
55. TR с дисконтом за номер предложения
56. На странице есть про 'оплату SMS'.
57. Магазинность страницы

58. Порнографичность страницы

59. Стихотворность документа

60. Максимальная стихотворность четверостишья

61. Язык документа - английский

62. Запрос полностью покрывается двумя точными группами, состоящими из exact match слов запроса подряд ((http://wiki.yandex-team.ru/poiskovajaplatforma/tr/CoverageByGroups Про покрытие группами))

63. Существует группа, состоящая из exact match слов запроса, покрывающая запрос (возможно, с пропуском, добавлением или заменой слова)

64. Доля запроса, покрываемая самой длинной группой, состоящей из любых хитов (в т.ч. словоформ и синонимов). Возможно, с пропуском, добавлением или заменой слова

65. Язык документа кириллический

66. Показывают насколько текст является неестественным с точки зрения русского языка. Оценка того, насколько можно считать текст документа сгенерированным синонимайзером либо вообще автоматическим. ((http://wiki.yandex-team.ru/JandeksPoisk/KachestvoPoiska/ObshayaFormula/TekushhieKomponenty/antispam?v=1il#h58953-2 подробнее))

67. Дата документа которая прописана на странице, ремапится квадратным корнем

68. У документа есть текстовая релевантность

69. BM25, где в качестве 'слов' выступают выделенные сегменты запроса

70. Вес' сегментов запроса в тексте

71. Показатель неестественности текста с точки зрения русского языка. Число плохих пар слов в тексте, перенормированное в отрезок [0,1] по формуле z/(z+10)

72. Доля плохих пар среди всех найденных в таблице: z/(x+1), где z – число плохих пар в тексте, а x – число ((http://wiki.yandex-team.ru/EvgenijjGrechnikov/TestSynonimizers 2000-релевантных)) пар

73. число латинских букв в тексте (не считая разметки), загнанное в [0,1] формулой n/(n+100)

74. Предыдущие факторы - исправленные

75. Число слов в тексте (Слово - то, что выделил леммер), отображается в [0,1] по формуле x/(x+A)

76. Число слов русского языка в заголовке

77. Средняя длина слова

78. Процент числа слов внутри тега <a>.. от числа всех слов

79. Процент числа слов вне тегов (вне скобок <>) от числа всех слов

80. Процент числа слов, являющихся 200 самыми частыми словами языка, от числа всех слов текста

81. Число использованных в тексте 500 самых популярных слов языка, деленное на 500

82. Логарифм среднего геометрического вероятностей триграмм в тексте. (вероятность триграммы - число ее встречаний в тексте, деленное на число всех триграмм) , отображается в [0,1] по формуле -x(x+A)

83. Логарифм среднего геометрического условных вероятностей триграмм. условная вероятность триграммы - ее вероятность, деленная на вероятность биграммы из первых двух слов

84. Разница между текущей датой и датой документа, определённой датировщиком, 1 - дата документа равна текущей, 0 - документу 10 лет или более, Если дата не определена, равен 0. Внимание!((1 - DaterAge)*60)^2 = возраст страницы в днях.

85. Максимальное число форм по всем словам запроса - max по всем словам запроса числа_форм_для_слова/64

86. Взвешенная по весам слов сумма числа форм - сумма по всем словам запроса числа_форм_для_слова/64*вес_слова; remap вида x/(1 + x).

87. Невзвешенная сумма числа форм - сумма по всем словам запроса числа_форм_для_слова/64/число_слов_запроса

88. Аналоги одноименных факторов, вес слова = 1

89. Доля разных частей речи в тексте. доля числительных (среди всех слов, у которых удалось распознать часть речи)

90. доля частиц

91. доля местоименных прилагательных

92. доля местоименных существительных

93. доля глаголов

94. доля слов, которые могут быть как существительными мужского рода, так и существительными женского рода, но не среднего рода, среди всех существительных (примеры: 'колибри' - пример неопределённого рода, который можно определять двумя способами, 'Александра' - омонимия).

95. Размер самого большого текстового сегмента страницы (из фактора [18] PureText)

96. DSSM модель с ранним связыванием, обученная на реформулировках и дообученная на ASR гипотезы музыкальных запросов к Алисе

97. DSSM модель с ранним связыванием, обученная на реформулировках и дообученная на музыкальные запросы к Алисе

98. Простой BM25 по парам слов - берем все пары слов запроса и считаем число их вхождений в текст документа. Вес =1. Комм Не работает, если в запросе есть стоп-слово

99. Язык документа соответствует языку запроса

100. На странице порно реклама

101. BM25 с разными параметрами для разных полей, включая входящий анкортекст. Веса текста входящих на страницу ссылок нормируются в зависимости от delta page rank ссылки

102. Фактор имени Buettcher, Clarke и Lushman (модифицированный) ((http://wiki.yandex-team.ru/JandeksPoisk/KachestvoPoiska/ObshayaFormula/TekushhieKomponenty/BCLm подробнее))

103. Униграммная языковая модель. Моделируется языковая по документу, сглаживается общеязыковой моделью. При построении модели по документу используется информацию о том, в каком поле документа встретилось слово запроса (Title, head или plain text)

104. Вычисляет покрытие запроса буквенными триграммами заголовка документа

105. Вычисляет покрытие заголовка буквенными триграммами заголовка документа

106. Считает сумму вхождений следующего вида: последовательность слов запроса длиной больше двух, встретившихся в одном предложении; нормировано на длину документа.

107. Оценивает минимальное расстояние между парами слов запроса с учетом удаленности пары от начала документа (Minimal Pair Size with Attenuation). Под парами понимаются все последовательные биграммы слов запроса. Таким образом, количество пар равно количеству слов в запросе, уменьшенному на 1. Соответственно, фактор имеет смысл для запросов, состоящих более чем из одного

слова.((http://wiki.yandex-team.ru/JandeksPoisk/KachestvoPoiska/ObshayaFormula/TekushhieKomponenty/MPSA MPSA))

108. Отличается от BCLm тем, что веса всех слов считаются одинаковыми. ((http://wiki.yandex-team.ru/JandeksPoisk/KachestvoPoiska/ObshayaFormula/TekushhieKomponenty/BCLm2 BCLm2))

109. Текстовая релевантность на основе языковой модели, учитывающая абсолютную позицию. Идем по тексту с окошком 20 слов, строим по каждому окошку языковую модель (то есть распределение вероятностей на словах русского языка) и вычисляем вероятность порождения запроса. За удаление от начала документа штрафуем модель.

110. Модификация фактора Bclm2, облегченная для использования в фастранке. Основное отличие состоит в том, что в BclmLite не используются абсолютные смещения слов относительно начала документа. Вместо этого фактор работает с обычными позициями вида <Номер_предложения, Позиция_в_предложении>. При этом близость между словами учитывается только внутри предложения.((http://wiki.yandex-team.ru/JandeksPoisk/KachestvoPoiska/ObshayaFormula/TekushhieKomponenty/BCLmLite BCLmLite))

111. Исправленный YmwFull. Отличается от предыдущей версии только поведением на 2хсловных запросах. ((http://wiki.yandex-team.ru/JandeksPoisk/KachestvoPoiska/ObshayaFormula/TekushhieKomponenty/YMW подробнее))

112. uses 'country aux tree' (auxqc)

113. Страница — '404' (доля токенов '404' по отношению к общему числу токенов на странице)

114. Фактор оценивает как слова запроса группируются друг с другом в тексте документа без учета их порядка. ((http://wiki.yandex-team.ru/SergejjKrylov/QueryWordCohesionTR описание))

115. Количество букв в сегменте Aux

116. Количество пробелов в сегменте Aux

117. Количество запятых в сегменте Content

118. Дисперсия IDF слов запроса при условии наличия текстовых хитов в документе (смешанный запросно-текстовый фактор)

119. Язык документа соответствует стране запроса

120. Доля сегментов запроса, присутствующая в тексте

121. Язык документа - один из допустимых для Турции (турецкий, английский, немецкий, французский, арабский, азербайджанский) либо документ имеет нулевую длину. На поисковой стадии вычисляется только для IsRealGeoLocal запросов.

122. Вариация на тему ((http://wiki.yandex-team.ru/JandeksPoisk/KachestvoPoiska/ObshayaFormula/TekushhieKomponenty/DBM25 DBM25)), см. ysite/yandex/relevance/dbm25.cpp

123. Популярность языка документа. Число от 0 до 1. ((http://wiki.yandex-team.ru/JandeksPoisk/KachestvoPoiska/ObshayaFormula/TekushhieKomponenty/LanguagePopularity LanguagePopularity))

124. Фактор оценивает отличия позиций слов в заголовке от позий слов в запросе

125. Размечается пул из PRS логов при помощи Bert, обученного на sinsig. На этом пуле обучается dssm модель, с использованием BaseRegionChain

126. BM25 заголовка страницы по её тексту

127. BM25 заголовка страницы по текстам ссылок на неё

128. Доля уникальных триграмм заголовка в триграммах ссылок

129. Доля уникальных триграмм ссылок в триграммах заголовка

130. Зарекламленность страницы

131. DBM отдельно по числам

132. DBM отдельно по гео-объектам запроса

133. DBM отдельно по существительным

134. Оценивает соответствие позиций слов в предложениях документа позициям слов в запросе.

135. В документе присутствует ФИО из запроса.

136. На документе есть прямая ссылка на файл

137. На документе есть ссылка на файлхостинг

138. Близость слов запроса к самому тяжелому слову.

139. Документ содержит пользовательский отзыв/комментарий

140. Функция правдоподобия распределения годов в документе. Временно отключен

141. Среднее арифметическое позиций дат в документе. Временно отключен

142. Доля слов документа из сегментов со score > 2.

143. Finetuned reformulations DSSM to commercial clicked bargain odd-like target from visit log

144. Фактор по ФИО из оригинального запроса Считается по содержимому документа. Алгоритм: Chain0Wcm

145. Запросно-документная модель навигационности.

146. Фактор по тексту запроса и заголовку (title) документа, оценка соответствия числовых диапазонов при словах-маркерах

147. Фактор по ФИО из оригинального запроса Считается по содержимому документа. Минимальный размер окна, в которой входят все слова запроса. Нормировано на число слов в запросе.

148. Фактор по ФИО из оригинального запроса Текст документа. Алгоритм CosineMatchMaxPrediction.

149. Фактор по всем ФИО из оригинального запроса Агрегация по всем расширениям. Тип аггрегации по расширениям: наибольшее значение фактора; Считается по содержимому документа. Алгоритм: Chain0Wcm

150. Фактор по всем ФИО из оригинального запроса Агрегация по всем расширениям. Тип аггрегации по расширениям: наибольшее значение фактора; Считается по содержимому документа. Минимальный размер окна, в которой входят все слова запроса. Нормировано на число слов в запросе.

151. Фактор по всем ФИО из оригинального запроса Агрегация по всем расширениям. Тип аггрегации по расширениям: наибольшее значение фактора; Текст документа. Алгоритм CosineMatchMaxPrediction.

152. DSSMное предсказание клика по данным, специфичным для Алисы

153. Фактор по телефонным аттрибутам tel_full из оригинального запроса Текст документа. Алгоритм агрегации весов слов Bocm15. Коэффициент нормализации 0.01.

154. Предсказание суммарного таймспента до конца сессии при условии реализации этой пары запрос-документ

155. Предсказание вклада этой пары запрос-документ в таймспент

156. Предсказание процента длины трека, который будет проигран при условии реализации этой пары запрос-трек

157. Bm15K01 factor over hits from Title

158. Bocm15K001 factor over hits from Title

159. Bm11Norm16384 factor over hits from Text

160. Bocm11Norm256 factor over hits from Text

161. CosineMatchMaxPrediction factor over hits from Text

162. Bm15FLogK0001 factor over hits from FieldSet2 stream

163. BclmWeightedFLogW0K0001 factor over hits from FieldSet3 stream

164. Bm15FLogW0K00001 factor over hits from FieldSetUT stream

165. Chain0Wcm factor over hits from Body

166. PairMinProximity factor over hits from Body

167. MinWindowSize factor over hits from Body

168. Нейронная модель качества контента для медицинской тематики

169. Нейронная модель качества контента для медицинской тематики (для экспов)

170. Нейронная модель качества контента для финансовой и юридической тематик

171. Нейронная модель качества контента для финансовой и юридической тематик (для экспов)

172. Фактор лингвистического бустинга. Тип расширений: RequestWithRegionName. Bm11 по тексту и тайтлу документа

173. Фактор лингвистического бустинга. Тип расширений: RequestWithRegionName. CosineMatchMaxPrediction по тексту и тайтлу документа

174. Фактор лингвистического бустинга. Тип расширений: RequestWithRegionName. Фактор: Bm15 по группе стримов 2.

175. Фактор лингвистического бустинга. Тип расширений: RequestWithRegionName. Фактор: BclmWeightedFLogW0 по группе стримов 3.

176. Фактор лингвистического бустинга. Тип расширений: RequestWithRegionName. Фактор Chain0Wcm по тексту документа

177. Нейронная модель качества контента для sos тематики

178. Нейронная модель качества контента для sos тематики (для экспов)

179. Предсказание таймспента сессии при условии реализации данной пары запрос-документ

180. Фактор по оригинальному запросу. Считается по заголовку документа. Алгоритм агрегации весов слов - BclmMixPlain: линейная смесь аннотационного Bclm веса и взвешенного Positionless веса слова, затем пословные счётчики агрегируются через bm15. Коэффициент нормализации 10^(-5).

181. Фактор по оригинальному запросу. Считается по заголовку документа. Алгоритм CMMatchTop5AvgMatchValue.

182. Фактор по оригинальному запросу. Считается по заголовку документа. Степень покрытия слов запроса с точностью до формы (без синонимов).

183. Фактор по оригинальному запросу. Считается по заголовку документа. Вес хита умножается на 1/ (1 + позиция слова в предложении) Алгоритм агрегации весов слов: Bm15. Коэффициент нормализации 0.5.

184. Фактор по оригинальному запросу. Считается по содержимому документа. Алгоритм агрегации весов слов - BclmMixPlain: линейная смесь аннотационного Bclm веса и взвешенного Positionless веса слова, затем пословные счётчики агрегируются через bm15. Коэффициент нормализации 10^(-5).

185. Фактор по оригинальному запросу. Считается по содержимому документа. Алгоритм CosineMatchMaxPrediction.

186. Фактор по оригинальному запросу. Считается по содержимому документа. Алгоритм AllWcmWeightedPrediction.

187. Фактор по оригинальному запросу. Считается по содержимому документа. Алгоритм агрегации весов слов Bocm15. Коэффициент нормализации 0.01.

188. Фактор по оригинальному запросу. Считается по содержимому документа. Алгоритм: QueryPartMatchSumValueAny.

189. Фактор по оригинальному запросу. Считается по содержимому документа. Степень покрытия слов запроса с точностью до формы (без синонимов).

190. Фактор по оригинальному запросу. Считается по содержимому документа. Степень покрытия слов запроса в точной форме.

191. Фактор по оригинальному запросу. Считается по содержимому документа. Алгоритм аггрегации весов: Bm15MaxAnnotation Коэффициент нормализации 0.01.

192. DSSM model trained on clicks. Takes bigrams into account. Embeddings for documents are computed offline.

193. Документная dssm модель language classifier rus.

194. Документная dssm модель language classifier eng.

195. Документная dssm модель language classifier other.

196. Предсказание DSSM модели для определения нерелевантных ответов Алисы

197. BM25FdPR с нормировкой на среднюю длину документа, зависящую от языка документа. Используются только хиты текстов.

198. DSSM model trained on click odd pool

199. DSSM model trained on click personalization pool

200. DSSM model trained on click triangle pool

201. Нейронная документная модель для поиска неожиданной жести

202. Исходный запрос с удалением глаголов. Считается по заголовку документа. Алгоритм агрегации весов слов: Bm15. Коэффициент нормализации 0.1.

203. Исходный запрос с удалением глаголов. Считается по компазиционном стриму, состоящего из токенезированного урла и заголовка документа. Алгоритм агрегации весов слов: Bm15FLogW0. Коэффициент нормализации 0.0001.

204. Исходный запрос с удалением глаголов. Считается по содержимому документа. Минимальный размер окна, в которой входят все слова запроса. Нормировано на число слов в запросе.

205. Фактор по фильтрованному оригинальному запросу: вычисляется dssm-расстояние от запроса без слов до исходного запроса, после чего происходит отсечение по порогу. Взвешенное объединение стримов Url,Title,Body,Links,CorrectedCtr,LongClick,OneClick,BrowserPageRank,SplitDwellTime, SamplePeriodDayFrc,SimpleClick,YabarVisits,YabarTime. Алгоритм агрегации весов слов: Bm15FLog (Bm15 агрегация логарифмов встречаемости слов). Коэффициент нормализации 0.001.

206. Фактор по фильтрованному оригинальному запросу: вычисляется dssm-расстояние от запроса без слов до исходного запроса, после чего происходит отсечение по порогу. Считается по компазиционном стриму, состоящего из токенезированного урла и заголовка документа. Алгоритм агрегации весов слов: Bm15FLogW0. Коэффициент нормализации 0.0001.

207. DSSM model trained on cross language CTRs using serp similarity hard miner.

208. Для всех слов слов запроса вычисляется вес методом query-mutation (расстояние между запросами при наличии и отсуствии слова). Берётся сумма весов слов найденных в тайтле, делённое на сумму весов всех слов.

209. Для всех слов слов запроса вычисляется вес методом query-mutation (расстояние между запросами при наличии и отсуствии слова). Берётся максимум веса среди слов, отсутсвующих в тайтле документа.

210. Результат применения нейронной модели, обученной отличать длинные клики от остальных событий, входом модели являтся пословные и биграмные счётчики, рассчитываемые по текстовым стримам (Body, Url).

211. Считается как (80-x) где x — возвраст документа в часах (непрерывно). Использует данные датировщика RobotAddTime

212. Считается как (10-x) где x — возвраст документа в днях (непрерывно). Использует данные датировщика RobotAddTime

213. Разница между текущей датой и датой документа, определённой датировщиком RobotAddTime, 1 — дата равна текущей, 0 — документу 10 дней и больше, или дата не определена

214. DSSM модель, которая предсказывает логарифм самого длинного клика на серпе. В качестве негативных примеров выбираем урлы из прошлых запросов этого же пользователя, причем максимальное время между запросами не более 7 минут (суперхарды по переформулировкам)

215. DSSM модель с ранним связыванием, обученная на переформулировках, которая предсказывает логарифм самого длинного клика на серпе.

216. Neural network value for contexts of query hits in document text. Predicts relevance-all-8-years. Uses formula ussr-dump-20190719 prs-20190720 all-8-years [t > 0.25] CrossEntropy 20k 0.25 -S 0.8 -Z 1 predictions for learning.

217. DSSM модель, обученная на пуле переформулировок, которая в запросной части помимо самого запроса получает 4 расширения XfDt с самым большим весом

218. Модель, обученная на предсказание оценки формулой ussr-dump-20190719 prs-20190720 all-8-years [t > 0.25] CrossEntropy 20k 0.25 -S 0.8 -Z 1.

219. Нейронная документная модель для поиска неожиданной жести (для экспов)

220. Модель, обученная на предсказание оценки формулой ussr-dump-20190719 prs-20190720 all-8-years [t > 0.25] CrossEntropy 20k 0.25 -S 0.8 -Z 1 и дообученная на оценки релевантности.