МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №11
по курсу «Информационный поиск»

«Сравнение качества работы поисковых моделей по мере NDCG»

Выполнил:
студент группы ИУ9-21М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2019

1 Цель работы

Необходимо сравнить следующие модели из Л.Р., выполненных ранее:

- Векторная, без idf.
- Векторная, TFIDF.
- Языковая.

При разметке релевантных документов использовать шкалу от 0 (предложение не содержит факта) до 2 баллов (содержит полный факт).

В Таблицах 1, 2, 3 приведены нагляные сравнения выдач моделей и разметка. Подписи таблиц – соответствующие запросы в нормализованном виде.

2 Ход работы

Необходимо оценить, какая система выдает лучший результат по мере $normDCG$.

$$normDCG = \frac{DCG}{idealDCG}, idealDCG = \sum^{Relevant} \frac{rel_i}{\log_2(i+1)}$$

где $DCG = \sum \frac{rel_i}{\log_2(i+1)}$

Посчитаем DCG для каждой модели и idealDCG для каждого запроса:

- В рецензии на компьютерную игру критик пожаловался на то, что смерть заставляет начинать уровень заново
  - $idealDCG = \frac{2}{\log_2 2} + \frac{1}{\log_2 3} = 2.63$ (факт в том или ином виде содержится в двух предложениях)
  - $DCG_{vect} = \frac{1}{\log_2 4} = 0.5, normDCG = \frac{0.5}{2.63} = 0.19$
  - $DCG_{tfidf} = \frac{1}{\log_2 3} = 0.63, normDCG = \frac{0.63}{2.63} = 0.24$
  - $DCG_{tfidf} = \frac{1}{\log_2 3} = 0.63, normDCG = \frac{0.63}{2.63} = 0.24$
- Под стенами осаждённой шведами русской крепости немцы побили шотландцев за пиво.
  - $idealDCG = \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{1}{\log_2 4} = 3.13$ (факт в том или ином виде содержится в трех предложениях)
  - $DCG_{vect} = \frac{1}{\log_2 3} + \frac{2}{\log_2 6} = 1.41, normDCG = 0.45$
  - $DCG_{tfidf} = \frac{1}{\log_2 2} + \frac{2}{\log_2 5} = 1.86, normDCG = 0.6$
  - $DCG_{lang} = \frac{1}{\log_2 3} = 0.63, normDCG = 0.2$
- Есть версия, что Джек Потрошитель был женщиной.

Таблица 1: рецензия компьютерный игра критик пожаловаться смерть заставлять начинать уровень заново

	Векторная модель
0	метр выпуск книга компьютерный игра быть указать игра хорошеть графика отличный звуковой сопровождение
0	заключение критик сообщить впечатлеть
1	потеря жизнь уровень запускаться заново жизнь последний игра заканчиваться
0	автор книга компьютерный мир посчитать игра отличный график очень неплохой музыка
0	разработать свой редактор игра позволять редактировать уровень график игра
	TFIDF
0	вскрывать брюшной полость джек потрошитель начинать уже смерть жертва
1	потеря жизнь уровень запускаться заново жизнь последний игра заканчиваться
0	критик отметить являться клон англ русск
0	заключение критик сообщить впечатлеть
0	метр выпуск книга компьютерный игра быть указать игра хорошеть графика отличный звуковой сопровождение
	Языковая модель
0	вскрывать брюшной полость джек потрошитель начинать уже смерть жертва
1	потеря жизнь уровень запускаться заново жизнь последний игра заканчиваться
0	метр выпуск книга компьютерный игра быть указать игра хорошеть графика отличный звуковой сопровождение
0	заключение критик сообщить впечатлеть
0	автор книга компьютерный мир посчитать игра отличный график очень неплохой музыка

  - $idealDCG = \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{1}{\log_2 4} = 3.13$ (факт в том или ином виде содержится в трех предложениях)
  - $DCG_{vect} = \frac{1}{\log_2 3} = 0.63, normDCG = 0.2$
  - $DCG_{tfidf} = \frac{2}{\log_2 2} + \frac{1}{\log_2 4} = 2.5, normDCG = 0.8$
  - $DCG_{lang} = \frac{2}{\log_2 2} = 2.0, normDCG = 0.63$

Усредним значения $normDCG$ для каждой модели:

- Векторная: $normDCG = \frac{0.19+0.45+0.2}{3} = 0.28$
- **TFIDF**: $normDCG = \frac{0.24+0.6+0.8}{3} = 0.55$
- Языковая: $normDCG = \frac{0.24+0.2+0.63}{3} = 0.36$

Модель TFIDF показала наилучший результат.

Наивная векторная модель и языковая модели показали примерно равные результаты, опередив друг друга в одном из запросов.

Таблица 2: стен осажда нной швед русской крепость немец побить шотландец пиво

	Векторная модель
0	неоднократный попытка швед совершить подкоп взорвать стена вовремя пресекаться защитник крепость
1	результат бойня погибнуть немец шотландец
0	шотландец бежать немец русский гарнизон везенберг быть поздний доставить москва
0	несколько год перемирие северный прибалтика вызвать русско литовский война год русский войско возобновить военный действие
2	март немец шотландец дело дошлый потасовка вызвать неоплаченный эль взаимный оскорбление
	TFIDF
1	результат бойня погибнуть немец шотландец
0	неоднократный попытка швед совершить подкоп взорвать стена вовремя пресекаться защитник крепость
0	шотландец бежать немец русский гарнизон везенберг быть поздний доставить москва
2	март немец шотландец дело дошлый потасовка вызвать неоплаченный эль взаимный оскорбление
0	камень алмаз останавливать свой падение оказываться земля стен
	Языковая модель
0	неоднократный попытка швед совершить подкоп взорвать стена вовремя пресекаться защитник крепость
1	результат бойня погибнуть немец шотландец
0	шведский финский солдат присоединиться шотландский преимущественно пехота немецкий преимущественно конница артиллерия наёмник
0	почти девятимесячный осада ревелеть оплот шведский владычество прибалтика увенчаться успех войско иван грозный удаться взять несколько меньший крепость частность вейсенштейн
0	шотландец бежать немец русский гарнизон везенберг быть поздний доставить москва

Таблица 3: версия джек потрошитель быть женщиной

	Векторная модель
0	утверждать джек потрошитель быть льюис кэрролл
1	возвращение джек потрошитель героиня фильм молли считать потомок известный убийца джек потрошитель присутствовать фильм качество персонаж виртуальный реальность собиратель душа сериал сезон серия который представляться версия тот джек потрошитель быть женщина
0	корнуэлла заявить джек потрошитель быть британский художник уолтер сикерта
0	потрошитель эпизод потрошитель сериал грань возможный возвращение джек потрошитель слэшер который присутствовать аллюзия способ убийство джек потрошитель
0	джек потрошитель перерезать горло слева направо рана быть очень глубокий
	TFIDF
2	основа быть взять женский версия убийца
0	придерживаться версия пять жертва
1	возвращение джек потрошитель героиня фильм молли считать потомок известный убийца джек потрошитель присутствовать фильм качество персонаж виртуальный реальность собиратель душа сериал сезон серия который представляться версия тот джек потрошитель быть женщина
0	один версия имя джек потрошитель скрываться душевнобольной польский еврей эмигрант аарон косминский
0	ставить этот версия сомнение однозначный доказательство тот жертва быть задушить существовать
	Языковая модель
2	основа быть взять женский версия убийца
0	придерживаться версия пять жертва
0	один версия имя джек потрошитель скрываться душевнобольной польский еврей эмигрант аарон косминский
0	ставить этот версия сомнение однозначный доказательство тот жертва быть задушить существовать
0	утверждать джек потрошитель быть льюис кэрролл

3 Выводы

В ходе работы были изучены сравнительные результаты работы трех моделей информационного поиска. Наилучший результат показала модель TFIDF.