

# Реализация моделей информационного поиска

Андрей Лебедев, группа 424

## 1 Постановка задачи

Целью данного задания является разработка системы для определения наиболее релевантных документов в корпусе по заданным запросам. Для ранжирования требуется реализовать 3 модели информационного поиска:

1. Модель частоты термов (TF).
2. Модель TF-IDF.
3. Языковая модель.

## 2 Ход решения

### 2.1 Подготовка корпуса

Для начала необходимо было подготовить текстовый корпус. Корпус документов был получен с помощью API Википедии. В качестве статей были выбраны те, что содержались в следующих фактах:

1. На московском острове растёт восьмиметровая облепиха.
2. Гватемальско-испанские отношения были разорваны после того, как посольство Испании было сожжено.
3. Булгаков не смог приехать на похороны матери из-за отсутствия денег.

Этапы подготовки корпуса:

1. **Сбор текстов.** Тексты были загружены через API Википедии по названиям статей.

2. **Разбиение на предложения.** Полученные тексты были разбиты на отдельные предложения с помощью библиотеки `razdel`.
3. **Удаление знаков препинания и стоп-слов.** Для каждого предложения удалялись все знаки препинания с помощью регулярного выражения. Затем из текста были удалены слова, которые не несут значимой смысловой нагрузки. Для этого был использован список русских стоп-слов из библиотеки `nlk`.
4. **Токенизация и нормализация.** Предложения были разбиты на токены с помощью библиотеки `razdel`, после чего каждый токен был приведён к нормальной форме с использованием морфологического анализатора `rumorphy2`. Это позволило обрабатывать разные формы одного и того же слова как одно и то же слово, что улучшает точность поиска.

На выходе данного этапа мы получили корпус, разбитый на документы (предложения), нормализованный и очищенный от ненужных символов и слов.

## 2.2 Модель частоты термов (TF)

Модель частоты термов (TF) оценивает релевантность предложения на основе того, сколько раз каждое слово из запроса встречается в предложении. Формально, частота терма  $tf(t, d)$  для терма  $t$  в документе  $d$  равна количеству вхождений терма в документ.

Для оценки релевантности используется косинусная мера сходства между вектором запроса и вектором документа. Каждый запрос и документ представляются в виде векторов, где каждое измерение соответствует терму, а его значение — частоте терма (TF). Косинусная мера вычисляется по формуле:

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

## 2.3 Модель TF-IDF

Модель TF-IDF улучшает TF за счёт добавления инверсной частоты документа, которая снижает вес часто встречающихся слов в корпусе:

$$\text{tfidf}(t, d) = \text{tf}(t, d) \times \log \frac{N}{\text{df}(t)},$$

где  $N$  — общее количество документов в корпусе, а  $\text{df}(t)$  — количество предложений, содержащих терм  $t$ .

## 2.4 Языковая модель

Языковая модель оценивает вероятность того, что данное предложение релевантно запросу. При тестировании начального решения для 1-го запроса была получена вероятность 0 для всех документов. Стало ясно, что проблема заключается в следующем: слово «восьмиметровый» достаточно редко, и не входит в корпус, поэтому сглаживания с помощью  $\lambda$  недостаточно. Так, один из множителей в итоговом произведении вероятности равен 0, и «зануляет» все выражение. Тогда было принято решение для расчёта вероятности использовать сглаживание Лапласа, которое добавляет небольшую ненулевую вероятность даже для слов, которые не встречаются в корпусе. Общая формула для языковой модели:

$$P(\text{sentence} \mid \text{query}) = \prod_{t \in \text{query}} (\lambda P_{\text{doc}}(t \mid \text{sentence}) + (1 - \lambda) P_{\text{corpus}}(t)),$$

где  $P_{\text{ml}}(t \mid \text{sentence})$  — вероятность термина  $t$  в предложении, а  $P_{\text{corpus}}(t)$  — вероятность термина в корпусе.

Формула для расчёта вероятности термина  $t$  в предложении с использованием сглаживания Лапласа выглядит следующим образом:

$$P_{\text{doc}}(t \mid \text{sentence}) = \frac{\text{count}(t, \text{sentence}) + \alpha}{|\text{sentence}| + \alpha \cdot V}$$

## 2.5 Модель FastText

В дополнение к ранее описанным моделям была реализована модель на основе векторов FastText. Для каждого предложения и каждого запроса были вычислены два типа векторов:

1. Усредненные векторы FastText.
2. Усредненные векторы FastText с учетом IDF.

Косинусное сходство между векторами запроса и предложений позволило получить ранжированные списки предложений для каждого запроса.

### 3 Результаты

Система возвращает ранжированные списки предложений из корпуса с использованием трёх моделей.

#### Запрос 1

«На московском острове растёт восьмиметровая облепиха»

##### Модель TF

Ранг 1, предложение 9, TF: 0.447214

На острове широко распространена облепиха.

Ранг 2, предложение 130, TF: 0.447214

Облепиха.

Ранг 3, предложение 70, TF: 0.282843

Орнитофауна острова типична для Москвы, например, на острове гнездится соловей.

##### Модель TF-IDF

Ранг 1, Предложение 130, TF-IDF: 0.367448

Облепиха.

Ранг 2, Предложение 9, TF-IDF: 0.247596

На острове широко распространена облепиха.

Ранг 3, Предложение 122, TF-IDF: 0.178568

— Облепиха иволистная — растёт на юге китайского автономного района Синьцзян, в горных районах Индийского субконтинента (Бутан, Индия, Непал).

## Языковая модель

Ранг 1, предложение 122, вероятность:  $1.249 * 10^{-16}$

— Облепиха иволистная — растёт на юге китайского автономного района Синьцзян, в горных районах Индийского субконтинента (Бутан, Индия, Непал).

Ранг 2, предложение 120, вероятность:  $1.243 * 10^{-16}$

== Классификация ==

Известны два вида:

*Hipporhaë rhamnoides* L. typus — Облепиха крушиновидная — растёт почти повсюду в Европе и в зоне умеренного климата Азии (встречается и в части тропической зоны — в Индии и Пакистане).

Ранг 3, предложение 109, вероятность:  $1.040 * 10^{-16}$

== Экология ==

Растут по берегам водоёмов, в поймах рек и ручьёв, на галечниках и песчаных почвах.

## Модель FastText с простым усреднением

Ранг 1, предложение 39, Сходство: 0.769358

После вывоза песка остров стал зарастать деревьями, и на нём появились бобры, давшие острову современное название.

Ранг 2, предложение 63, Сходство: 0.765195

Если в Покровском-Стрешневе бобры выступают в роли условных вредителей, то на острове они повреждают только малоценный самосев деревьев и при этом изменяют биотоп острова, являясь средообразующим видом (бобры рубят деревья и кустарники и строят хатки).

Ранг 3, предложение 117, Сходство: 0.760899

Кустарники облепихи используются для закрепления песков, дорожных откосов, оврагов, для живых изгородей.

## Модель FastText с усреднением по IDF

Ранг 1, предложение 117, Сходство: 0.735771

Кустарники облепихи используются для закрепления песков, дорожных откосов, оврагов, для живых изгородей.

Ранг 2, предложение 45, Сходство: 0.732635

Широко распространена облепиха (отдельные её экземпляры достигают высоты 7—8 метров), часто встречается рейнутрия сахалинская.

Ранг 3, предложение 36, Сходство: 0.731894

На Бобровом была также проведена подсыпка грунта за счёт углубления русла реки и использования строительного мусора, включая разбитые гранитные плиты, применявшиеся в облицовке набережной, из которого были насыпаны холмы высотой до пяти метров.

## Запрос 2

«Гватемальско-испанские отношения были разорваны после того, как посольство Испании было сожжено.»

## Модель TF

Ранг 1, предложение 132, TF: 0.516398

— 192 с. Гватемальско-испанские отношения — двусторонние дипломатические отношения между Гватемалой и Испанией.

Ранг 2, предложение 192, TF: 0.408248

Испания имеет посольство в Гватемале.

Ранг 3, предложение 198, TF: 0.408248

В результате случившегося Испания на 4 года разорвала дипломатические отношения с Гватемалой.

## Модель TF-IDF

Ранг 1, предложение 132, TF-IDF: 0.467785

— 192 с.

Гватемальско-испанские отношения — двусторонние дипломатические отношения между Гватемалой и Испанией.

Ранг 2, предложение 198, TF-IDF: 0.377055

В результате случившегося Испания на 4 года разорвала дипломатические отношения с Гватемалой.

Ранг 3, предложение 168, TF-IDF: 0.322530

2 февраля 1980 года Испания разорвала дипломатические отношения с Гватемалой из-за инцидента в посольстве и угроз в адрес дипломатического персонала.

## Языковая модель

Ранг 1, предложение 132, вероятность:  $2.612 * 10^{-19}$

— 192 с.

Гватемальско-испанские отношения — двусторонние дипломатические отношения между Гватемалой и Испанией.

Ранг 2, предложение 168, вероятность:  $2.332 * 10^{-19}$

2 февраля 1980 года Испания разорвала дипломатические отношения с Гватемалой из-за инцидента в посольстве и угроз в адрес дипломатического персонала.

Ранг 3, предложение 242, вероятность:  $2.322 * 10^{-19}$

Как только испанский посол покинул воздушное пространство страны пребывания, 2 февраля 1980 года Испания разорвала дипломатические отношения с Гватемалой из-за инцидента в посольстве и угроз в адрес дипломатов.

## Модель FastText с простым усреднением

Ранг 1, предложение 242, Сходство: 0.827318

Как только испанский посол покинул воздушное пространство страны пребывания, 2 февраля 1980 года Испания разорвала дипломатические отношения с Гватемалой из-за инцидента в посольстве и угроз в адрес дипломатов.

Ранг 2, предложение 247, Сходство: 0.820830

=== Последствия === Правительство Гватемалы выступило с заявлением, в котором утверждалось, что его силы якобы вошли в посольство по просьбе испанского посла и что оккупировавшие посольство, которых они называли «террористами», «принесли в жертву заложников, а затем сожгли себя».

Ранг 3, предложение 157, Сходство: 0.817998

Протестующие решили войти в посольство Испании, поскольку по их мнению эта страна им сочувствовала.

## Модель FastText с усреднением по IDF

Ранг 1, предложение 247, Сходство: 0.798225

=== Последствия === Правительство Гватемалы выступило с заявлением, в котором утверждалось, что его силы якобы вошли в посольство по просьбе испанского посла и что оккупировавшие посольство, которых они называли «террористами», «принесли в жертву заложников, а затем сожгли себя».

Ранг 2, предложение 198, Сходство: 0.798010

В результате случившегося Испания на 4 года разорвала дипломатические отношения с Гватемалой.

Ранг 3, предложение 242, Сходство: 0.777889

Как только испанский посол покинул воздушное пространство страны пребывания, 2 февраля 1980 года Испания разорвала дипломатические отношения с Гватемалой из-за инцидента в посольстве и угроз в адрес дипломатов.



## Запрос 3

«Булгаков не смог приехать на похороны матери из-за отсутствия денег.»

### Модель TF

Модель TF: Ранг 1, предложение 618, TF: 0.782624

Михаил Булгаков не смог приехать на похороны из Москвы в Киев из-за отсутствия денег.

Ранг 2, предложение 64, TF: 0.267261

Бобры на острове не образуют плотин из-за отсутствия водотоков.

Ранг 3, предложение 432, TF: 0.250000

А Булгаков — незаконное!»

### Модель TF-IDF

Ранг 1, предложение 618, TF-IDF: 0.866028

Михаил Булгаков не смог приехать на похороны из Москвы в Киев из-за отсутствия денег.

Ранг 2, предложение 64, TF-IDF: 0.282097

Бобры на острове не образуют плотин из-за отсутствия водотоков.

Ранг 3, предложение 287, TF-IDF: 0.188777

Если ему хотелось проехаться на такси на последние деньги, он без раздумья решался на этот шаг: «Мать ругала за легкомыслие.

### Языковая модель

Ранг 1, предложение 618, вероятность:  $1.251 * 10^{-25}$

Михаил Булгаков не смог приехать на похороны из Москвы в Киев из-за отсутствия денег.

Ранг 2, предложение 287, вероятность:  $2.952 * 10^{-26}$

Если ему хотелось проехать на такси на последние деньги, он без раздумья решался на этот шаг: «Мать ругала за легкомыслие.

Ранг 3, предложение 64, вероятность:  $2.906 * 10^{-26}$

Бобры на острове не образуют плотин из-за отсутствия водотоков.

### **Модель FastText с простым усреднением**

Ранг 1, предложение 618, Сходство: 0.901601

Михаил Булгаков не смог приехать на похороны из Москвы в Киев из-за отсутствия денег.

Ранг 2, предложение 281, Сходство: 0.799727

Это можно увидеть в воспоминаниях Татьяны Николаевны: «Фаты у меня, конечно, никакой не было, подвенечного платья тоже — я куда-то дела все деньги, которые отец прислал.

Ранг 3, предложение 287, Сходство: 0.790209

Если ему хотелось проехать на такси на последние деньги, он без раздумья решался на этот шаг: «Мать ругала за легкомыслие.

### **Модель FastText с усреднением по IDF**

Ранг 1, предложение 618, Сходство: 0.934395

Михаил Булгаков не смог приехать на похороны из Москвы в Киев из-за отсутствия денег.

Ранг 2, предложение 281, Сходство: 0.793424

Это можно увидеть в воспоминаниях Татьяны Николаевны: «Фаты у меня, конечно, никакой не было, подвенечного платья тоже — я куда-то дела все деньги, которые отец прислал.

Ранг 3, предложение 287, Сходство: 0.785442

Если ему хотелось проехать на такси на последние деньги, он без раздумья решался на этот шаг: «Мать ругала за легкомыслие.

## 4 Выводы

В ходе выполнения работы были реализованы четыре модели для ранжирования предложений по их релевантности запросам: модель частоты термов (TF), модель TF-IDF, языковая модель и FastText.

1. TF полезна для быстрого и простого подсчёта встречаемости ключевых слов, но даёт худшие результаты из-за недостатка понимания контекста.
2. TF-IDF предоставляет более точные результаты, поскольку снижает вес общих слов и акцентирует внимание на редких и значимых словах.
3. Языковая модель обладает более гибким подходом к ранжированию, учитывая вероятность слов как в предложении, так и в корпусе. Тем не менее, на маленьком корпусе ей не удалось проявить себя сильно лучше других подходов.
4. Добавление модели FastText показало, что использование предобученных векторов и IDF-взвешивания позволяет получить более точные результаты. FastText помог лучше представить семантические связи между запросами и предложениями.