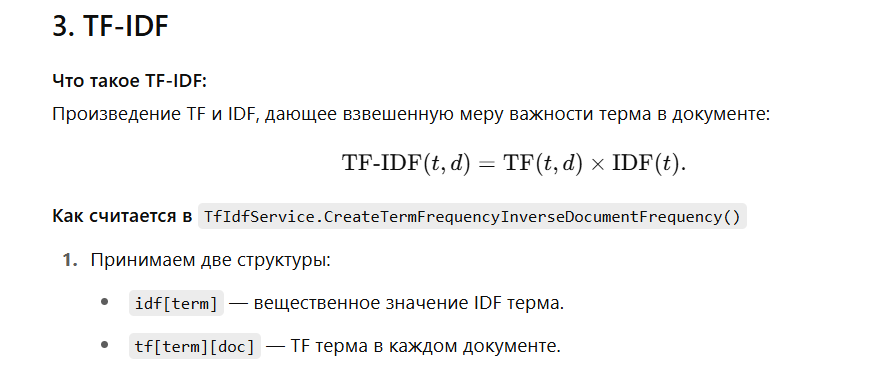
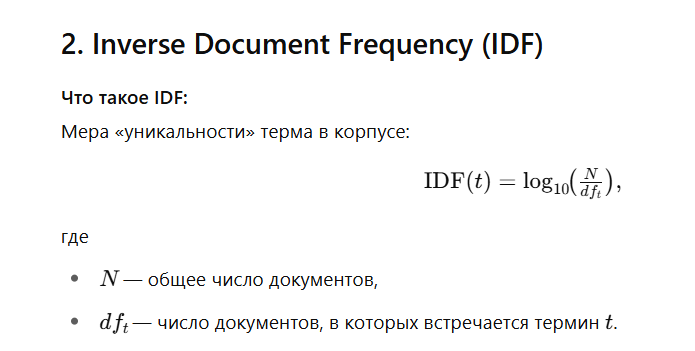
4 дз  
  
 **TF** показывает, как часто термин встречается внутри одного документа.

 **IDF** показывает, насколько термин редок во всём наборе документов.

 **TF‑IDF** объединяет оба показателя: частые в конкретном документе и редкие в корпусе получают наивысший вес.





Подумав пару секунд

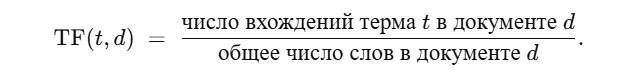
Ниже — максимально подробное разборное описание каждого компонента, их задач и примеры интерпретации полученных чисел.

**1. TF (Term Frequency)**

**1.1. Что это такое**

TF — это просто **частота** терма в одном документе, нормированная на размер этого документа. Позволяет понять, насколько часто слово встречается внутри данного текста.

**1.2. Формула**

**1.3. Для чего нужен**

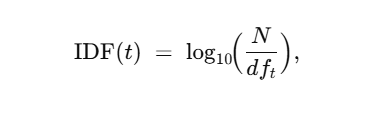
* Помогает оценить «важность» терма **в конкретном документе**: чем чаще слово встречается в одном тексте, тем более значимо оно для его содержания.
* Без нормировки (просто raw count) тяжело сравнивать документы разной длины: у длинного документа всегда будет больше «сырых» вхождений всех слов. TF нивелирует этот эффект.

**2. IDF (Inverse Document Frequency)**

**2.1. Что это такое**

IDF отражает, насколько **редок** терм по всему корпусу документов. Чем более распространён термин, тем ниже его информационная ценность (например, «и», «в», «на»).

**2.2. Формула**

где

* N — общее число документов в коллекции,
* dftdf\_t — число документов, где термин tt встречается хотя бы один раз.

Округляют обычно до 5–6 знаков после запятой.

**2.3. Для чего нужен**

* Уменьшает вес **часто встречающихся** слов и повышает вес **редких**, понижая шум от служебных слов («стоп-слов»).
* Дает «обратную» меру важности: редкий термин (низкий dftdf\_t) → высокий idfidf → значимое отличие в документах.

**2.4. Пример**

Пусть корпус состоит из N=4N = 4 документов:

| **Термин** | **Документы с термом** | **dftdf\_t** | **IDF(t)\mathrm{IDF}(t)** |
| --- | --- | --- | --- |
| «кот» | d1, d2 | 2 | log⁡10(4/2)=0,301030\log\_{10}(4/2)=0{,}301030 |
| «мышь» | d2 | 1 | log⁡10(4/1)=0,602060\log\_{10}(4/1)=0{,}602060 |
| «и» | d1, d2, d3, d4 | 4 | log⁡10(4/4)=0\log\_{10}(4/4)=0 |

**Интерпретация**:

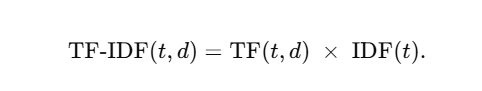
* «и» встречается в каждом документе → idf=0idf=0 (никакой различительной силы).
* «кот» встречается в половине документов → умеренный idf≈0,301idf≈0{,}301.
* «мышь» — только в одном → более высокий idf≈0,602idf≈0{,}602.

**3. TF–IDF**

**3.1. Что это такое**

TF–IDF — взвешенная мера, сочетающая локальную частоту терма в документе и его глобальную редкость.

**3.2. Формула**

Округляют до тех же 5–6 знаков.

**3.3. Для чего нужен**

* Подчёркивает термы, которые **одновременно** часто встречаются в данном документе и при этом редки во всём корпусе.
* Классически применяется для:
  + Векторного представления документов (векторный поиск, классификация).
  + Выделения «ключевых слов» текста.
  + Фильтрации нерелевантных (слова с idf≈0idf≈0).

**3.4. Пример сквозного расчёта**

Допустим, для документа d1d\_1 из примера выше:

* TF(«кот»,d1)=0,025\mathrm{TF}(\text{«кот»}, d\_1)=0{,}025, IDF(«кот»)=0,301030\mathrm{IDF}(\text{«кот»})=0{,}301030
* TF(«мышь»,d1)=0,010\mathrm{TF}(\text{«мышь»}, d\_1)=0{,}010, IDF(«мышь»)=0,602060\mathrm{IDF}(\text{«мышь»})=0{,}602060

Тогда:

* TF-IDF(«кот»,d1)=0,025×0,301030≈0,007526\mathrm{TF\text{-}IDF}(\text{«кот»}, d\_1) = 0{,}025 \times 0{,}301030 ≈ 0{,}007526
* TF-IDF(«мышь»,d1)=0,010×0,602060≈0,006021\mathrm{TF\text{-}IDF}(\text{«мышь»}, d\_1) = 0{,}010 \times 0{,}602060 ≈ 0{,}006021

**Анализ**:

* Хотя «кот» встречается чаще, чем «мышь» (TF выше), «мышь» чуть более редкое слово (IDF выше), поэтому их TF–IDF-значения получаются сопоставимыми.
* Любой стоп-слово («и») даст TF–IDF = TF × 0 = 0 — полностью отфильтровано.

**4. Практическая интерпретация значений**

1. **Большой TF–IDF**  
   →\rightarrow терм важен для этого документа и помогает отличить его от других.
2. **Низкий (но ненулевой) TF–IDF**  
   →\rightarrow термин встречается в документе, но либо он редок (низкий TF), либо слишком частый по корпусу (низкий IDF).
3. **TF–IDF = 0**  
   – либо TF = 0 (терм не встречается в этом документе),  
   – либо IDF = 0 (терм встречается во всех документах и не несёт различительной информации).

**Пример анализа для нескольких документов**

| **Термин** | **TF(d1)** | **IDF** | **TF–IDF(d1)** | **TF(d2)** | **TF–IDF(d2)** |
| --- | --- | --- | --- | --- | --- |
| «кот» | 0.025 | 0.3010 | 0.00753 | 0.015 | 0.00452 |
| «мышь» | 0.010 | 0.6021 | 0.00602 | 0.000 | 0.00000 |
| «собака» | 0.005 | 0.4771 | 0.00239 | 0.020 | 0.00954 |

* Для **d1** самым информативным термином будет «кот» (0.00753 против 0.00602 и 0.00239).
* В **d2** «собака» (0.00954) важнее «кот» (0.00452), «мышь» не встречается совсем → 0.

**5. Выводы**

* **TF** измеряет локальную важность слова в документе.
* **IDF** измеряет глобальную редкость слова в корпусе.
* **TF–IDF** сочетает оба показателя, чтобы выделить термы, значимые **именно** для данного документа.
* По **векторному пространству** TF–IDF легко оценивать сходство документов и запросов (через косинусную меру) и ранжировать результаты поиска.

Hw 5

