

Task

Overview

You are working in a team of software engineers, data scientists, developers and yourself. The team is

dedicated to push the boundaries in the area of information retrieval, but needs help with obtaining a

baseline to compare their results against. You have been assigned to implement this baseline method.

Baseline method

A popular task in information retrieval is to find a document d in a collection of documents D (known as a

'corpus') that is most relevant to a query string s .

A basic problem in retrieval is to measure how relevant a word is to a document in a corpus. The following sections describe one possible such measure, given a word w and a document d in a corpus D .

Word Importance (wi)

This value measures how important a word is in a document. While there are a number of ways in practice to calculate the word importance, for the purposes of this project, assume the importance of a

word w in a document d is given by:

$$wi(w, d) = f_{w, d} / M$$

where:

- $f_{w, d}$ is the frequency of w in d
- M is the total number of words in d

Generality Discount (gd)

Some words occur in natural language much more frequently than others. For example, the word 'is' will

occur much more frequently than the word 'magical'. The gd value seeks to inversely weigh a word based

on how frequently it occurs in a corpus. For this project, the gd value of a word w in a corpus D is given as

follows.

$$gd(w, D) = \log N/n$$

where

- N is the number of documents in corpus
- n is the number of documents that contain w

Word Relevance (wr)

After calculating the word importance and generality discount, the word relevance value is calculated as

the element-wise matrix multiplication of the wi and gd values.

$$wr(w, d, D) = wi(w, d) * gd(w, D)$$

Example

Consider the following table, which lists the counts of some words in a corpus consisting of 3 documents.

Word/Counts	Document1	Document2	Document3
This	10	12	5
is	8	5	4
dog	2	3	0
magical	0	0	1

The total words in each document are:

Document1	Document2	Document3
200	400	25

Given the above corpus and a search string "**This is magical**", the wr calculations for document 1 are as

follows.

Term	$wi(w, d_1)$	$gd(w, D)$	$wr(w, d_1)$
"this"	$10/200 = 0.05$	$\log 3/3 = 0$	$0.05 * 0 = 0$
"is"	$8/200 = 0.04$	$\log 3/3 = 0$	$0.04 * 0 = 0$
"magical"	$0/200 = 0$	$\log 3/1 = 0.477$	$0 * 0.477 = 0$

The Challenge

Given the following query strings and the 20 documents included under the "documents" directory,

design and implement a system to calculate the wr values. You must use **Python** and you may use any

Python library you prefer.

Query Strings

1. "tennis match"
2. "88 thousand people!"
3. "the plastic container; see "

Hints

- Only plain text, alpha-numeric characters are useful for analysis. You may filter out any punctuation or markup.
- If you think there is information missing, you may use any resource (except enlisting the help of others) and your own judgement to make assumptions about the problem.
- Please document your assumptions and design decisions. We will discuss these during the in-person interview.