Information Retrieval: Page ranking report

CS50 Final project

Parham Afsharnia

Introduction

Information Retrieval (IR) can be defined as a software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories, particularly textual information.

Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature i.e. usually text which satisfies an information need from within large collections which is stored on computers. For example, Information Retrieval can be when a user enters a query into the system. It can be classified in NLP science.[1]

Part 1.1

Tokenizer function removes any non-alphabetic characters, any kinds of whitespace and symbols.
Tokenizer is a custom-implemented function.
Numbers are not removed from the text, also alphabet letters were considered as tokens.

Part 1.2

a. Vocabulary size on this data collection : 8149
Vocabulary size means number of unique terms in a dataset or text word collection.

b. 10 most common word(s):
('the', 20204), ('of', 14032), ('and', 7116), ('a', 6835),
('in', 5034), ('to', 4725), ('is', 4118), ('for', 3714),
('with', 2444), ('are', 2431)

Obviously, the most common and frequent words and terms in human writings
Are units like ‘a’, ‘the’, ‘is’, ‘or’, ‘and’ and so on.
In NLP, these words are called “stop words”.

c. meaningful:
None of the words above are meaningful individually. stop words are extremely common in almost every text, although we can see the stop words below.

[1] Information retrieval

d. minimum number of words: 92
# this number represent the most common words that contain half of the text
This number is smaller than the minimum number of dataset without considering stopwords . Again because there are a lot.

Part 1.3

a. 5355 is vocabulary size after removing stop words and stemming
b.
c. 10 most common word(s):
('flow', 2184), ('W', 1402),('A', 1401), ('B', 1401),('I', 1400), ('T', 1400),('boundary', 1373),
('pressure', 1331),('layer', 1192), ('number', 1033)

After removing stopwords 10 common words in the dataset are changed and there are some meaningful words and meaningless words in there.
This information is important to us, because most commonly used words can give useful information about the content.

d. meaningful:
('flow', 2184), ('boundary', 1373), ('pressure', 1331),
('layer', 1192), ('number', 1033)
meaningless:
[('W', 1402), ('A', 1401), ('B', 1401), ('I', 1400), ('T', 1400)]

Part 1.4

The Heap’s low:

Heaps’ law describes the portion of a vocabulary which is represented by an instance document (or set of instance documents) consisting of words chosen from the vocabulary. This can be formulated as

Heaps’ law means that as more instance text is gathered, there will be diminishing returns in terms of discovery of the full vocabulary from which the distinct terms are drawn.[2]

[2] Heap's law

Part 2.1

1. tf/idf weighting scheme, implemented and document rankings for each query

TF-IDF (term frequency-inverse document frequency)

is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.

This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

It has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP).

TF-IDF was invented for document search and information retrieval. It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they don't mean much to that document in particular.[3]

Part 2.2

a. For each query in queries lists, a ranked list of documents determined it sthey are shown in pairs (queryID,docID) with the scoring numbers.
b. K Top most related information can be shown in the first page of the search engine.

[3] TF/IDF

acknowledgement

[1]. An Introduction to Information Retrieval, By Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze,chapter *5 Index compression,page 88*