Information Retrieval

Final Project Report

Tech Mania

Introduction

This project is mainly about searching about the Latest Tech news and articles about them. For this project,I crawled articles about them from web which is the data for us.Based on the query given by the user the top 10 documents were retrieved using a ranked retrieval algorithm

Ranked Retrieval algorithm : **TF-IDF**.

Documents retrieved based on Boolean retrieval algorithm.

Task 1: Scraping, Indexing, Ranked Retrieval

- Dataset for this retrieval system is crawled from: https://www.reuters.com/news/archive/technologynews
- Retrieved around 30,000 documents

1	web-scrap	web-scrap	title	title-href	descriptior	dates
2	16079443:	https://wv	Accor step	https://wv	Accor said	Jun 03 2015
3	16079452:	https://wv	Uber push	https://wv	Transporta	Jan 11 2016
4	16079523:	https://wv	Red Cross	https://wv	The Red Cı	May 26 2020
5	16079459:	https://wv	Zuckerberg	https://wv	Facebook	Jun 16 2016
6	16079514:	https://wv	Amazon dı	https://wv	Amazon.c	Oct 29 2019
7	16079476:	https://wv	With 'stick	https://wv	In the wor	Jul 05 2017

- All the terms which are extracted are preprocessed using different preprocessing steps
- Then, Stored the inverted index using Trie data structure because of faster retrieval of the docs corresponding to terms and then reated max tf for all the documents retrieved
- Then Implemented Ranked Retrieval for all the documents
- When we enter the query it will compute the tf-idf for the query terms and it will retrieve the query term present in the docs and we calculate max-tf of the term in each docs

```
Enter a query: amazon sales
```

- And then by multiplying the max-tf of document and tf-idf of the query we get scores for the each document corresponding to the query term and we have to take the top 10 scores by this we can retrieve the docs that are similar to the query

```
==========================================

doc ID =  1607950184-20673
title score =  1.3328584809576718

https://www.reuters.com/article/us-amazon-com-results/amazon-sales-outlook-falls-short-after-record-holiday-quarter-idUSKCN1PP2XK

Amazon.com Inc on Thursday forecast first-quarter sales below Wall Street estimates, warning that new regulations in India had created uncertainty around one of its key growth markets and saying it would step up investments in 2019.  ...

tf-idf score= 16.719030181844033


==========================================

doc ID =  1607947744-12360
title score =  1.3328584809576718

https://www.reuters.com/article/us-amazon-com-results/amazon-plows-ahead-with-high-sales-and-spending-profit-plunges-idUSKBN1AC339

Amazon.com Inc on Thursday reported a jump in retail sales along with a profit slump, as its rapid, costly expansion into new shopping categories and countries showed no sign of slowing.  ...

tf-idf score= 16.719030181844033


==========================================
```

Task 2: Document Similarity

- Dataset for this Document similarity is crawled from: https://www.reuters.com/news/archive/technologynews (Above Dataset only)
- Now we will check the similarity of documents with the help of pretrained doc2vec model
- We will give 2 input :
  - 1: Doc id (as a string) i.e. to compare similarity
  - 2: Doc id (List of strings) i.e. to be compared with

```
source_doc = 'how to delete an invoice'
target_docs = ['delete a invoice', 'how do i remove an invoice', 'purge an invoice']
```

- We will get output as a score :

```
[ {'score': 0.99999994, 'doc': 'delete a invoice'},
{'score': 0.79869318, 'doc': 'how do i remove an invoice'},
{'score': 0.71488398, 'doc': 'purge an invoice'} ]
```

Note: Above pics are examples of document similarity algorithm