Practice II

Document similarity

Introduction

- Access to scientific information is essential for research and technology development
- There are multiple repositories that allow you to search and consult scientific articles
- Publications have two access schemes: closed and open. In the closed scheme, a subscription must be paid, while in the open scheme, access is free for all users

Scientific article repositories

- ArXiv. A free distribution service and an open-access archive for scholarly articles in the following fields:
    - Physics
    - Mathematics
    - Computer Science
    - Biology
    - Finace
    - Electrical engineering
    - Economics
- Papers on arXiv are not peer-reviewed
- https://arxiv.org/

Scientific article repositories

- PubMed. Is the National Library of Medicine’s free, searchable bibliographic database supporting scientific and medical research
- It does not include full text journal articles access
- Papers on PubMed are peer-reviewed
- https://pubmed.ncbi.nlm.nih.gov/

Objective

- Develop a program that allows searching and retrieving articles collected from the arXiv and PubMed repositories using a text query

Specifications

- In team of 3-4 members do the following activities:

  1) Collection of articles from the selected repositories using *Web Scraping*

  2) Text normalization using *Spacy or NLTK*

  3) Text representation in a vector space model using *scikit-learn*

  4) Retrieval of the most similar articles to a query using *Cosine Similarity*

Collection of articles

- From the arXiv repository, articles from the following sections should be downloaded:
    - *Computation and Language*
    - *Computer Vision and Pattern Recognition*
- From the PubMed repository, articles from the *trending section* should be downloaded
- The collection for both repositories will be carried out until 300 items from each repository are completed
- In the case of arXiv, 150 articles from each section must be downloaded

ArXiv articles

- The content to be obtained from the arXiv articles is as follows:
  - DOI
  - Title
  - Authors
  - Abstract
  - Section
  - Publication date
- The articles are available in up to 3 different formats. To facilitate content extraction, the HTML format should be used

PubMed articles

- The content to be obtained from the PubMed articles is as follows:
    - DOI
    - Title
    - Authors
    - Abstract
    - Journal name
    - Publication date
- PubMed has a refence format (similar to RIS) from which content can be obtained more easily

ArXiv raw corpus

- The article’s content from arXiv should be saved in a corpus with the following format

DOI	Title	Authors	Abstract	Section	Date
10.48550./arXiv.<id_1>	<Title_1>	<Author_1_1, Author_2_1, …, Author_n_1>	<Abstract_content_1>	<Section_A>	<dd/mm/yyyy>
...	...	...	...	...	...
10.48550./arXiv.<id_m>	<Title_m>	<Author_1_m, Author_2_m, …, Author_n_m>	<Abstract_content_m>	<Section_B>	<dd/mm/yyyy>

- The corpus must be saved in a file named *arxiv_raw_corpus.csv* using tab character as field separator

PubMed raw corpus

- The article’s content from PubMed should be saved in a corpus with the following format

DOI	Title	Authors	Abstract	Journal	Date
DOI_1	<Title_1>	<Author_1_1, Author_2_1, …, Author_n_1>	<Abstract_content_1>	<Journal_name_1>	<dd/mm/yyyy>
...	...	...	...	...	...
DOI_m	<Title_m>	<Author_1_m, Author_2_m, …, Author_n_m>	<Abstract_content_m>	<Journal_name_n>	<dd/mm/yyyy>

- The corpus must be saved in a file named *pubmed_raw_corpus.csv* using tab character as field separator

Text normalization

- Apply the following normalization process to the fields Title and Abstract of the raw data corpora:
  - Tokenization
  - Remove stop words from the following grammatical categories: articles, prepositions, conjunctions, and pronouns
  - Lemmatization
- For stop words use POS tagging process to identify grammatical category
- The normalized version of corpora should be saved in csv files, with the same format a the previous ones, called *arxiv_normalized_corpus.csv* and *pubmed_normalized_corpus.csv*

Text representation

- Generate frequency, binarized and TF-IDF vector representations of the columns:
    - Title
    - Abstract
- The characteristics to be extracted are:
    - Unigram
    - Bigrams
- The resulting representations should be saved in *pkl* files

Retrieval of the most similar articles

- The query provided for this task is a BibTeX or RIS file
- Once the file has been selected, the following must be indicated
  - The comparative content of the article (Title or Abstract)
  - The features to be extracted (Unigram or Bigram)
  - The type of vector representation (frequency, binary or TF-IDF)
- Do the following with this document:
  - Apply the same normalization process performed to the normalized corpus
  - Extract the specified features
  - Generate the indicated vector representation
  - Apply the cosine similarity algorithm to determine the similarity between the input document and the rest of the documents in both corpus using the comparative content
  - Display the 10 most similar documents in descending order

Interface

- An interface must be created for the three main tasks:
  - Article collection
  - Text normalization and representation
  - Retrieval of similar articles

Article collection interface

- The interface for collecting articles should allow specifying the repository from which the articles are to be downloaded (arXiv or PubMed)
- The expected output is the raw corpus of articles collected with the specified format

Text normalization and representation interface

- This interface should allow specifying the raw corpus
- The expected outputs are the normalized corpus (1 per repository) and the pkl files of the text representation (12 files per repository)

Article retrieval interface

- This interface should allow specifying the following:
    - The file to be used as query (BibTeX of RIS)
    - The comparative content of the article (Title or Abstract)
    - The features to be extracted (Unigram or Bigram)
    - The type of vector representation (frequency, binary or TF-IDF)
- The expected output is the list of the 10 most similar documents in descending order

Evidence

- Source code
- Generated raw corpus
- Generated normalized corpus
- Pkl files of the text representation
- Document in PDF with the following table:

Test document <test_num>	**<test_text>**			
Corpus article	*Vector representation*	*Extracted features*	*Comparison content*	*Similarity value*

- Where:
  - <test_num>: number of the test file (1, 2, 3, ...)
  - <test_text>: content of the test file
- The document must include the names of the team's members
- All the members must upload the evidence

Evidence

Test document 1	Understanding the Limits of Lifelong Knowledge Editing in LLMs			
Corpus document	*Vector representation*	*Extracted features*	*Comparison element*	*Similarity value*
55	TF-IDF	Unigrams	Title	0.65
60	TF-IDF	Unigrams	Title	0.60
120	TF-IDF	Unigrams	Title	0.55
10	TF-IDF	Unigrams	Title	0.50
100	TF-IDF	Unigrams	Title	0.48
45	TF-IDF	Unigrams	Title	0.4
230	TF-IDF	Unigrams	Title	0.38
70	TF-IDF	Unigrams	Title	0.30
12	TF-IDF	Unigrams	Title	0.25
134	TF-IDF	Unigrams	Title	0.18