Multimedia Retrieval Project Report

Realized by: Eya Baklouti

Level: Master's degree in computer science

Semester: Winter Semester 2021/2022

Contents

1. Introduction

1.1. Project Scope

Writers or Bloggers often write articles, texts or a story that they don't have an illustration for, and some websites, newspapers require providing a photo, a cover image that is related to the topic in order be able to put the article according to the webpage, magazine design template.
The idea of the project is to create a tool that takes a text (the article, blog text) extracts the keywords from it, and based on those keywords it displays photos that are in the context or related to the topic discussed in the text.

1.2. Project specifications

The solution is a text-based image search engine which :

1) Preprocess a text
2) Extract a set of the most important words .
3) Use the set keywords to extract set images that are related to those keywords.

2. Solution workflow

The processes starts with inputing a text, the first step is to process the text through count vectorization to get the keyword candidates, filtering then perform word embeddings to be able to get an embedding of the word candidate , then we have to calculate the similarity between the keywords and add diversity to finally get a list of the most relevant keywords that are going to be used to select the most relevant images from the database and this by embedding them with the images embeddings and find the most semantically relevant images for the keywords of the imputed text

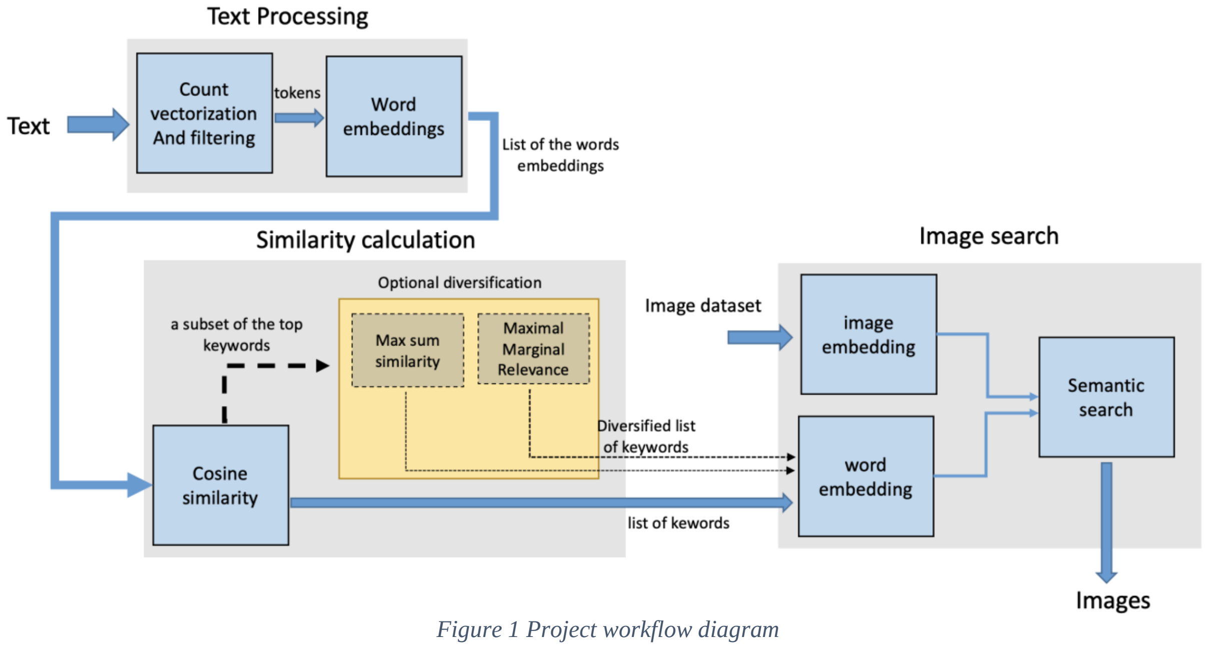


Figure 1 Project workflow diagram

3. Programming enviroment

3.1. Framework and Libraries :

- Sentence transformers
- Pandas
- Streamlit
- Sklearn
- Nltk
- Numpy
- Pytorch

3.2. Software enviroment :

- Anaconda
- Visual Studio code

4. Solution implementation

4.1 Text processing and Key word extraction

4.1.1. Candidate keywords/keyphrases :

- ***Ngram***
  In a document, N-grams are continuous sequences of words, symbols, or tokens. They can be defined as the adjacent sequences of items in a document in technical terms. In this project i choose to use the Ngram range of (1,1) basically because one word is so sufficient to be used for the image search .
- ***Stop words***
  We eliminate the stop words that have no specific meaning in the text and are frequently used , each language has its own stop words so since my project is destinated for the english texts i used the stop word list related to the english language
- ***Count vectorizer***
  I used the CountVectorizer function that scikit-learn library provides which allowed me to transform a text to a vector based on the frequency of each word in the text.

4.1.2. Word Embeddings :

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation.

BERT for word embedding:

Bidirectional Encoder Representations from Transformers (BERT) is a Transformer based machine learning methodology created by Google for NLP pre-training. It can be used to extract high-quality language elements from text data.

The BERT pre-trained model that i used for word embeddings is ***distilbert-base-nli-mean-tokens***
It is a sentence-transformer model and it maps sentences and paragraphs to a 768 dimentional dense vector space and can be used for tasks like clustering or semantic search.

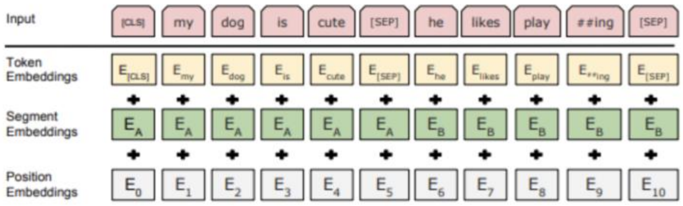


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

BERT, like most deep learning models focused at solving NLP-related problems, transforms each input token into a vector representation using a Token Embedding layer. BERT features additional embedding layers in the form of Segment Embeddings and Position Embeddings, unlike other deep learning models.

4.1.3 Similarity calculation

Cosine similarity

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.
It is often used to measure document similarity in text analysis which is the use case of the project.

4.1.3 Diversitfication

To avoid getting a list of key words with similar meanings, I included two functions in my solution to diversify the extracted key words, which would strike a careful balance between the accuracy of keywords/keywords and their diversity.
The algorithms used to create the functions are as follows:

Maximum sum similarity :

The maximum sum distance between pairs of data is defined as the pairs of data for which the distance between them is maximized. In the instance of my project, I want to increase candidate similarity to the document while minimize candidate similarity.

Maximal Marginal Relevance

Maximal Marginal Relevance (MMR). In text summarizing tasks, MMR aims to reduce repetition and increase diversity of findings. Fortunately, a keyword extraction algorithm called EmbedRank has implemented a version of MMR that give the possiblity to use it for diversifying our keywords/keyphrases.

4.2 Image search

4.2.1. Image dataset :

The image dataset used in this project is provided by the **sbert.net** website and it is accesibale through this link http://sbert.net/datasets/
It is a 1.8 Go dataset that containes 24996 image related to a lot of general and different topics which suites well my project since there is no specific topic or domain for the images because the texts are going to be general, diversified, and random.

4.2.2. Embedding Model

Image embeddings

For image embedding and for fulfilling the requirement of the project I choose a model that can perform image and word embedding for the keywords selected into a joint vector space.
The model is provided by the sentence-Transformers library, and it is used mainly for finding similar images and for image search.
The model name is **clip-ViT-B-32** and CLIP stands for Contrastive Language-Image Pre-Training which is a trained neural network on a variety of (image, text) tuples, the model can predict the most relevant image given a text, without any required optimization for the task, it additionally give the same performance of the original ResNet50 on ImageNet “zero-shot” without using any of the original 1.28M labeled examples.

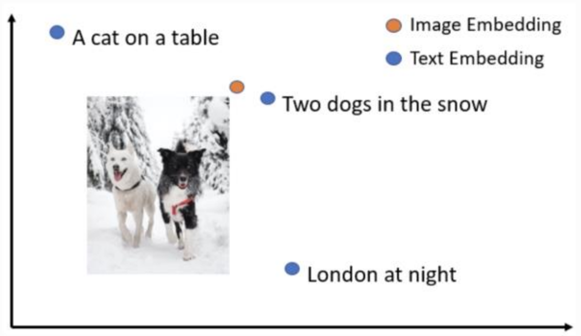
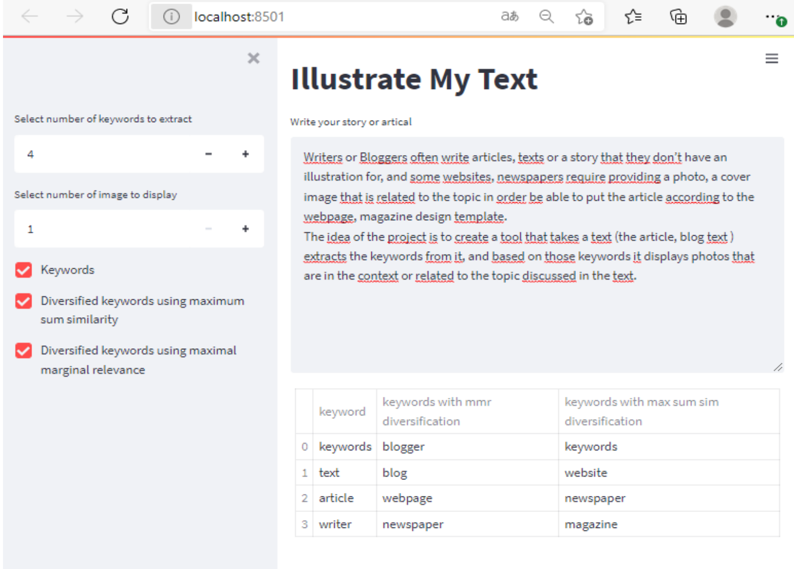


Figure 2 Illustration of image and text embedding

5. Results and perspective:

To better present the results I developed a simple web interface using streamlit library the figure below shows a screenshot of the interface. The interface allows the user to control the number of keywords that he wanted to extract as well as the number of photos that he wants to display and the possibility to select how much diversified the key words and the photos that he wat to get.
In the screenshot example I putted as a text the general description idea of my project in order to get some photos that I can put them as a cover photo to the web interface or if I wanted to write a short blog about the project.



Illustrate My Text

Select number of keywords to extract

4

Select number of image to display

1

☑ Keywords

☑ Diversified keywords using maximum sum similarity

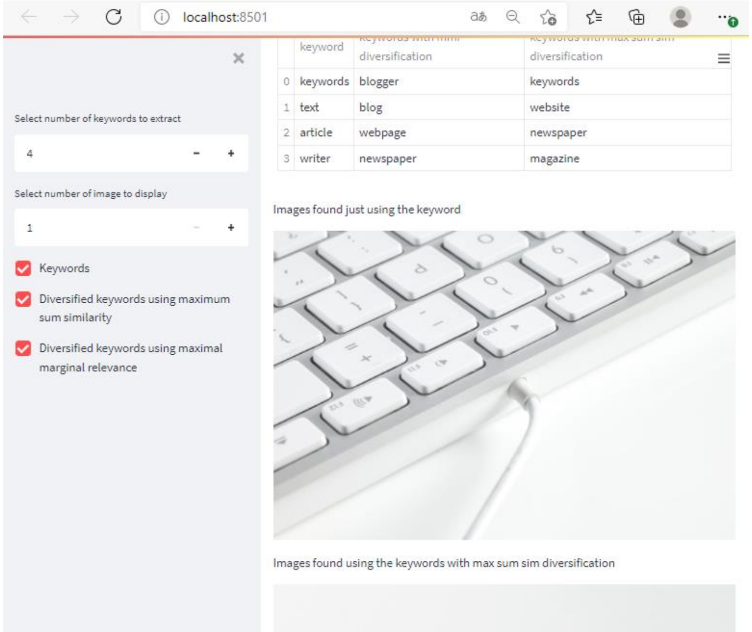
☑ Diversified keywords using maximal marginal relevance

Write your story or artical

Writers or Bloggers often write articles, texts or a story that they don’t have an illustration for, and some websites, newspapers require providing a photo, a cover image that is related to the topic in order be able to put the article according to the webpage, magazine design template.
The idea of the project is to create a tool that takes a text (the article, blog text) extracts the keywords from it, and based on those keywords it displays photos that are in the context or related to the topic discussed in the text.

	keyword	keywords with mmr diversification	keywords with max sum sim diversification
0	keywords	blogger	keywords
1	text	blog	website
2	article	webpage	newspaper
3	writer	newspaper	magazine

Figure 3 screenshot demo for the solution results 1



	keyword	keywords with mmr diversification	keywords with max sum sim diversification
0	keywords	blogger	keywords
1	text	blog	website
2	article	webpage	newspaper
3	writer	newspaper	magazine

Select number of keywords to extract: 4

Select number of image to display: 1

☑ Keywords

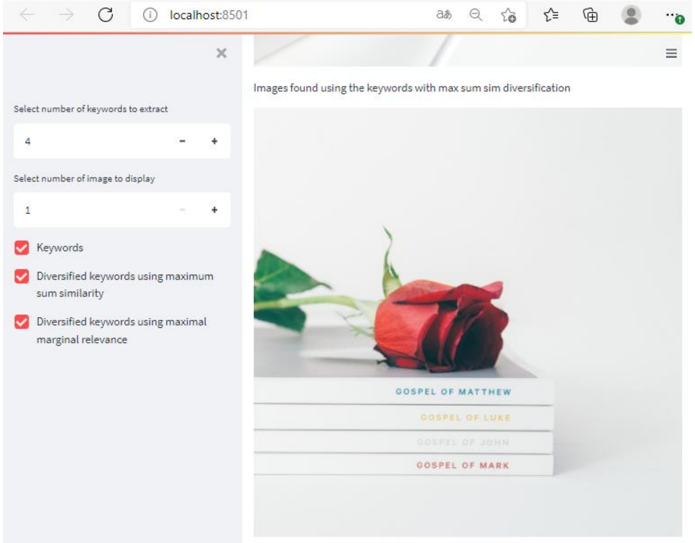
☑ Diversified keywords using maximum sum similarity

☑ Diversified keywords using maximal marginal relevance

Images found just using the keyword

Images found using the keywords with max sum sim diversification

Figure 4 screenshot demo of the solution result 2



Select number of keywords to extract: 4

Select number of image to display: 1

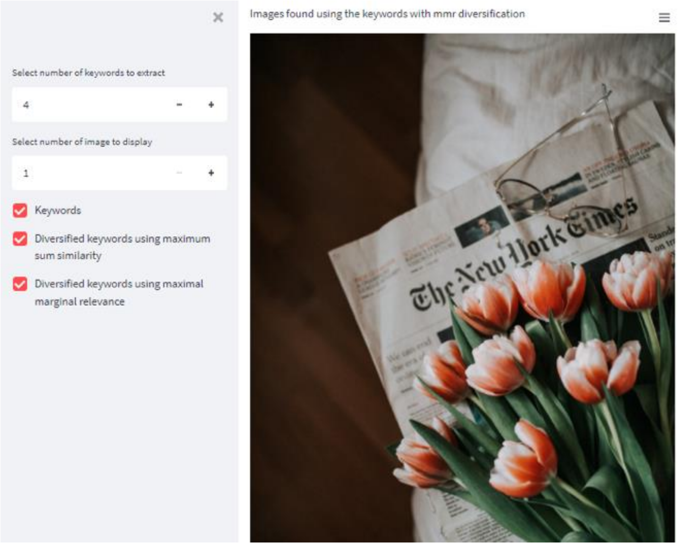
☑ Keywords

☑ Diversified keywords using maximum sum similarity

☑ Diversified keywords using maximal marginal relevance

Images found using the keywords with max sum sim diversification

Figure 5 screenshot demo of the solution results



Select number of keywords to extract: 4

Select number of image to display: 1

☑ Keywords

☑ Diversified keywords using maximum sum similarity

☑ Diversified keywords using maximal marginal relevance

Images found using the keywords with mmr diversification

Figure 6 screenshot demo f the solution results

Yet there is a lot to improve in this project, there is a lot of features and functionalities that can be added like clustering the text to which domain it belongs in addition to adding more flexibility in the parameters like adding the possibility of changing the ngram range as well as adding more images to the used dataset to get better results.

6. References

Papers

Nils Reimers andIryna Gurevych. **Sentence-bert: Sentence embeddings using siamese bert-networks**. CoRR, abs/1908.10084, 2019.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter** arXiv preprint arXiv:1910.01108 .

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever **Learning Transferable Visual Models From Natural Language Supervision**

websites

Sentence Transformers documentation https://pypi.org/project/sentence-transformers/
Hugging face documentation https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens
https://www.sbert.net/