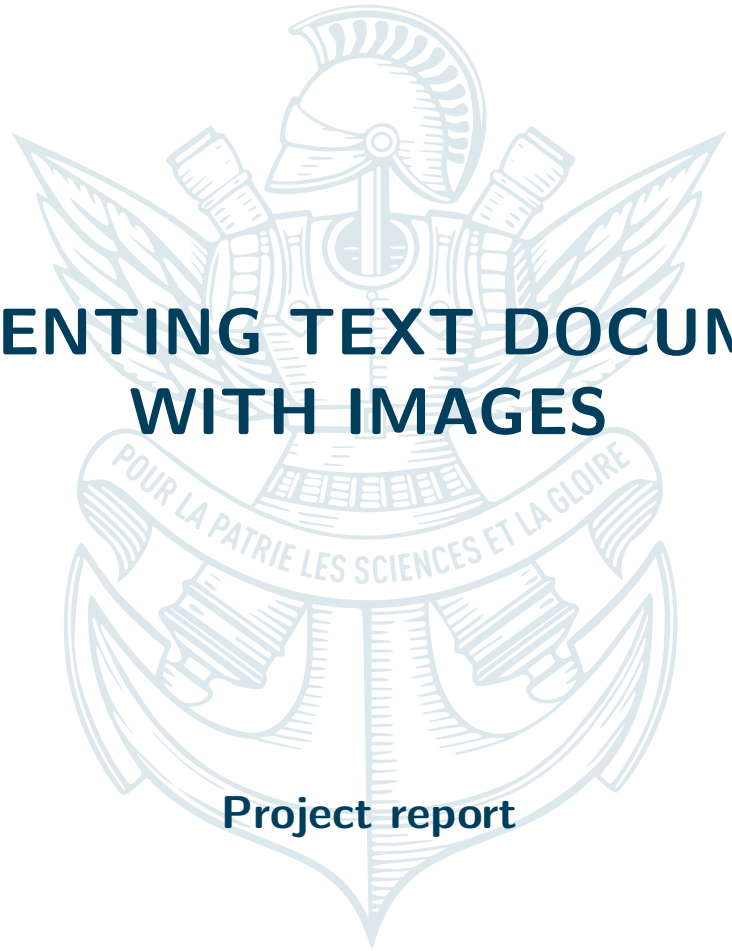

AUGMENTING TEXT DOCUMENTS WITH IMAGES

Project report

19 décembre 2017

Maxime Bourliatoux

1 PROBLEM STATEMENT

1.1 Global problem

Imagine that you are a journalist, a blogger or an editor. After you write an article in text, you would like to insert images from a large image database such as ImageNet to make the article interesting to the reader and catch his attention. Given the variety of articles that an editor may write, a manual process to retrieve such images is tedious and time-consuming.

In this project, we aim to design a software tool with intelligent algorithms for automating the processing of augmenting text documents with images. We have two specific goals :

1. The images retrieved must match the topic of a given article that an editor is editing.
2. Often times retrieving images only based on the topic of the document is not sufficient. The editor is likely to have his own interpretation of whether an image is relevant or not.

The project will be composed of two parts :

1. Design an algorithm that, given a text, retrieves images form the image database relevant to the theme of the article.
2. Implement an "Explore-by-Example" approach to have the closest match possible to the author's idea.

1.2 Part 1 - Fetch relevant images

The first part of the problem is about fetching images from a database corresponding to the theme of an article. To do so, we have two things to do : First we need to get the topic of the article. Then having the theme we have to find images related in the image database. Those two tasks will need two different algorithms, we will do one after the other and then combine the results. The objective here is to have a base algorithm so that at least fetched images will not be irrelevant.

1.3 Part 2 - Implement "Explore-by-Example" approach

In the second part of the project, we will try to improve the algorithm, in order to make it more useful. Indeed the first part of the project could be replaced by a search on Google Images for the main topic and does not reflect the intent of the author. Therefore we will try to implement an "Explore-by-Example" approach, to let the author guide the algorithm in its images choices.

Therefore the algorithm should first recommend a few images based on the topic, and solicit the user feedback on the images. Then the algorithm should incorporate such feedback to adjust the internal model of the editor's interest, and retrieve a few more images for feedback. This process goes in iterations. We would like to design an algorithm that requires a minimum number of iterations to return a specified number of images that the editor deems relevant and interesting.

2
LITERATURE SURVEY

2.1 Finding the theme of the article

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). *Automatic keyword extraction from individual documents.* Text Mining : Applications and Theory, 1-20.
`https://pdfs.semanticscholar.org/5a58/00deb6461b3d022c8465e5286908de9f8d4e.pdf`
This article presents the Rapid Automatic Keyword Extraction (RAKE) algorithm. An unsupervised method for extracting keywords from individual documents.

Singhal, A. (2001). *Modern information retrieval : A brief overview.* IEEE Data Eng. Bull., 24(4), 35-43.

Madhu Kumari, Akshat Jain, Ankit Bhatia, *Synonyms Based Term Weighting Scheme : An Extension to TF.IDF*, In Procedia Computer Science, Volume 89, 2016, Pages 555-561, ISSN 1877-0509
`http://www.sciencedirect.com/science/article/pii/S1877050916311589`
This article explore the possibility to add synonyms recognition to the TF*IDF.

Khoo Khyou Bun, Mitsuru Ishizuka (2002) *Topic Extraction from News Archive Using TF*PDF Algorithm*
`http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1181645&tag=1`
This article presents an improvement of the TF*IDF algorithm called TF*PDF. It avoids missing the topic of the text if this topic is in multiple articles.

Kewen Chen, Zuping Zhang, Jun Long, Hao Zhang, *Turning from TF-IDF to TF-IGM for term weighting in text classification*, In Expert Systems with Applications, Volume 66, 2016, Pages 245-260, ISSN 0957-4174
`http://www.sciencedirect.com/science/article/pii/S0957417416304870`
This article presents a variation of the TF*IDF algorithm that is supposed to outperform it.

2.2 Retrieving images in the database

To retrieve images based on the ImageNet database, I plan on using WordNet with the python package NLTK. Indeed, the ImageNet database is organized according to the WordNet hierarchy. So with a base word, using NLTK we can find a corresponding "Synset" corresponding to an image set in the database.

3
TECHNICAL CONTRIBUTIONS

3.1 Context of the experiments

To conduct the experiments, I will work with texts and images covering two topics : athletics and water areas such as oceans, seas, lake. Those two topics are different enough so that it will be quite easy to tell if the algorithm has made mistakes. They are also both quite documented, with a lot of texts and images on those topics. I have selected several texts from Wikipedia covering those subjects and a group of images.

Texts

`https://en.wikipedia.org/wiki/Marathon`
`https://en.wikipedia.org/wiki/100_metres`
`https://en.wikipedia.org/wiki/4_%C3%97_100_metres_relay`
`https://en.wikipedia.org/wiki/110_metres_hurdles`

`https://en.wikipedia.org/wiki/Ocean`
`https://en.wikipedia.org/wiki/Sea`
`https://en.wikipedia.org/wiki/Lake`
`https://en.wikipedia.org/wiki/Pond`

Images

I downloaded several images from the ImageNet database (`www.image-net.org`) and organized them in the same way than the database. I took images from the target clusters.

3.2 First results : RAKE

For the first version of the algorithm I used RAKE as it is an unsupervised algorithm, working with a single file. It allows us to have first results.

Results :

Marathon	100m	Ocean
olympic marathon	world record	ocean
boston marathon	race	planet
world record	women	earth
world records	break	seas
marathon distance	time	surface
marathon running	set	water
training	barrier	life
races		oceans
record		titan
distance		world

We see that the main idea behind the articles is here, but seems not enough to find good illustration images then.

This results are promising because they show that we can find easily a first idea of the topic. We now need to implement an algorithm based on the TF*IDF to try to improve the results.

4 TIME-LINE AND MILESTONES

— December : Creation of the development environment, first experiments, implementation of the RAKE algorithm.
— End of December & January : Implementation of the algorithm fetching an image based on the topic found.
— Then : Improvement of the part 1 of the project (implementation of TF-IDF) and search for articles to implement the part 2.