DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Operations/Image processing: Automatic defect detection

Topic modeling on the abstracts of research papers with the keyword “automatic defect detection”

Academic course: Business and Project Management

Student: Carmine Tranfa

1. Introduction

Point of view: a water utility company that would like to understand if it's possible to reduce the operation costs on the assets and infrastructures using automatic defects detection.

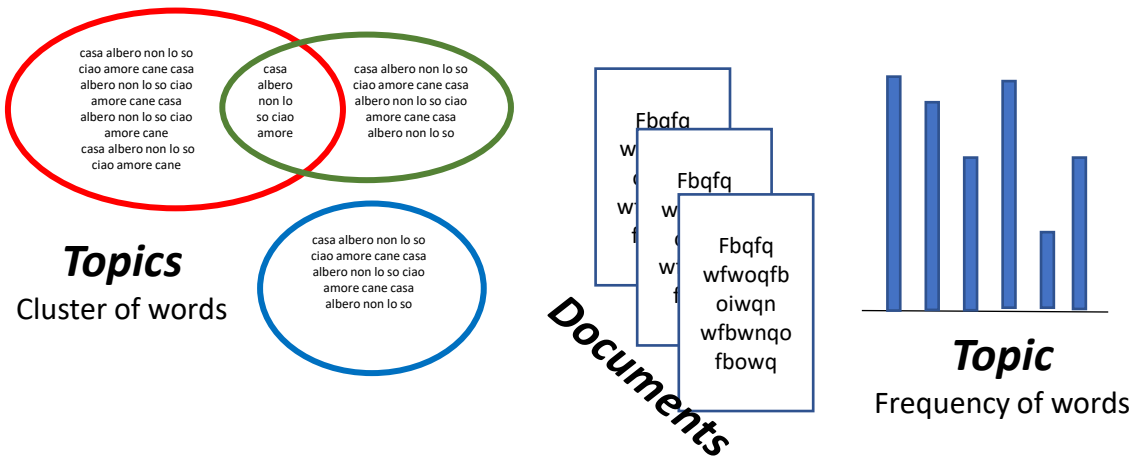
Project's aim: analyze the scientific literature about the topic, in order to understand in which industries this tool is used most and if there is some experiences about its use in the water utilities field.

2. Process

2.1 Scrape the data

2.2 Check and clean the data

2.3 Topic modeling



2.1 Scrape the data

Web scraping papers' information in two steps

- download a **paper's list** from Google Scholar which contains the keywords

"*automatic*" "*defect*" "*detection*"

	Paper Title	Year	Author	Publication	Url of paper
0	Fabric defect detection and classification usi...	1995.0	YF Zhang, RR Bresee	journals.sagepub.com	https://journals.sagepub.com/doi/abs/10.1177/0...
1	Approaches for improvement of the X-ray image ...	2019.0	W Du, H Shen, J Fu, G Zhang, Q He	Elsevier	https://www.sciencedirect.com/science/article/...
2	The application of one-class classifier based ...	2017.0	M Zhang, J Wu, H Lin, P Yuan, Y Song	Elsevier	https://www.sciencedirect.com/science/article/...

- scrape all the **papers' abstracts** from the links related to the articles.

Conventional image analysis hardware was used to image solid-shade, unpattened, woven fabrics. Two different software approaches for detecting and classifying knot and slub defects were studied and compared. The approaches were based on either gray level statistics or morphological operations. The autocorrelation function was used for both methods to identify fabric structural repeat units, and statistical or morphological computations were based on these units. Plain weave and twill weave fabrics were used to compare the performance of each software approach.

2.2 Check and clean the data



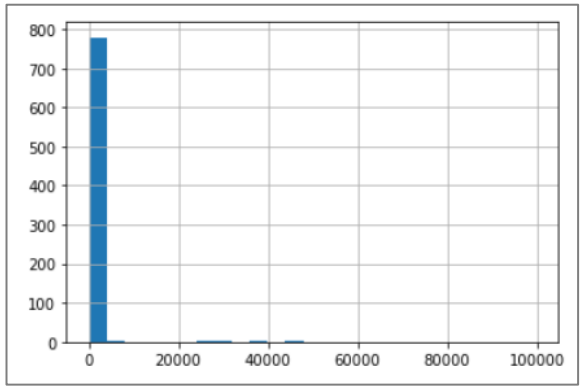
Loop activities

- Check for abstracts' average length distribution
- Find the length‘s outlier threshold “T”
- Abstracts’ analysis with more than “T” words
- Clean the abstracts from the wrong’s words scraped
- Repeat until the average abstract’s length distribution isn’t like a Gaussian

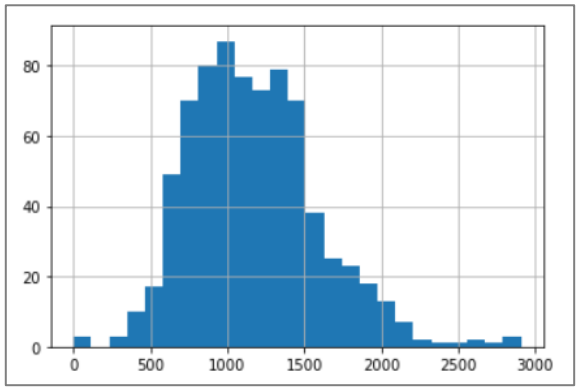
Abstracts visual analysis

```
ID: 469 - https://www.scientific.net/AMR.734-737.2898
ID: 489 - https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.638
ID: 497 - AbstractDetection of substation equipment can prom ...
ID: 498 -



Click on “Download PDF” for the PDF version o ...
ID: 500 - https://search.ieice.org/bin/summary.php?id=e90-c_11_21
ID: 522 - https://ascelibrary.org/doi/abs/10.1061/(ASCE)CP.1943-5
ID: 524 - https://www.matec-conferences.org/articles/matecconf/ab
0.html
ID: 526 - https://search.proquest.com/openview/0f9b53f5a0712b6f4c
ID: 527 - https://www.scientific.net/AMR.295-297.1274
ID: 534 - AbstractAutomatic defect detection is an important ...
ID: 543 - AbstractIn recent years, more and more scholars de ...
ID: 552 - https://search.proquest.com/openview/446c4242a1d7c4b471
```

Abstracts' length analysis



Abstracts' length output



2.3 Topic modeling - *definition*

What is Topic modeling?

Topic modeling is a method for unsupervised classification of documents, similar to clustering on numeric data, which finds some natural groups of items (topics) even when we're not sure what we're looking for.

Why to use it?

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives. It can help with the following:

- discovering the hidden themes in the collection;
- classifying the documents into the discovered themes;
- using the classification to organize/summarize/search the documents.

Which model use?

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

It is one of the most popular topic modeling methods. Each document is made up of various words, and each topic also has various words belonging to it. The aim of LDA is to find topics a document belongs to, based on the words in it.

Source: https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

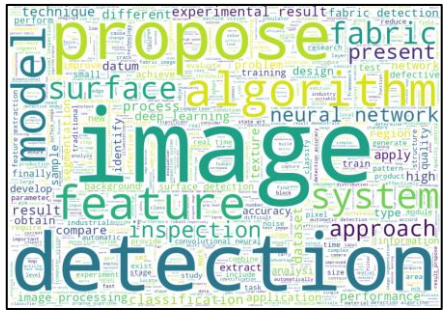
2.3 Topic modeling - *process*

Data pre-processing

- Lowercasing;
- Tokenization(Gensim);
- Punctuations and “Stop Words” removal (SpaCy);
- Lemmatization (SpaCy “en_core_web_lg” model).

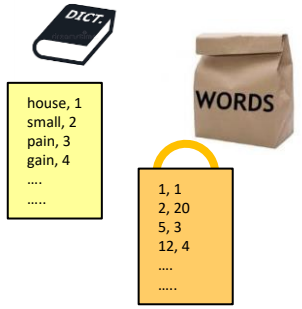
Exploratory Analysis

- Preview;
- Dummy word removal.



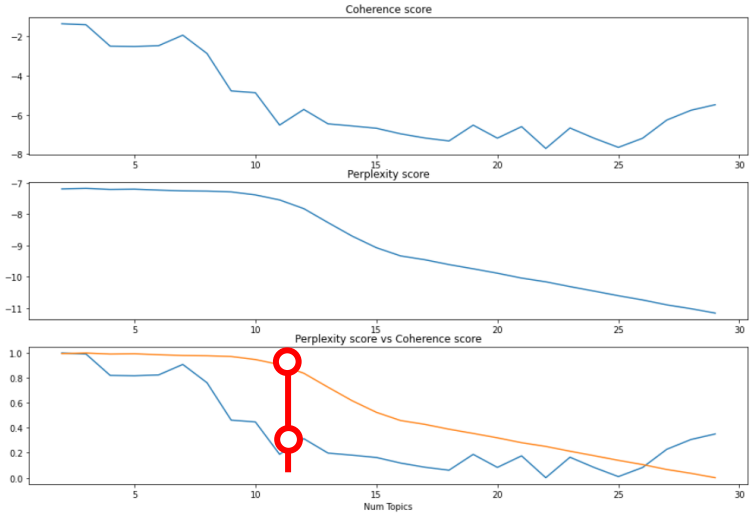
Data preparation

- Dictionary;
- Bag-of-words or corpus.

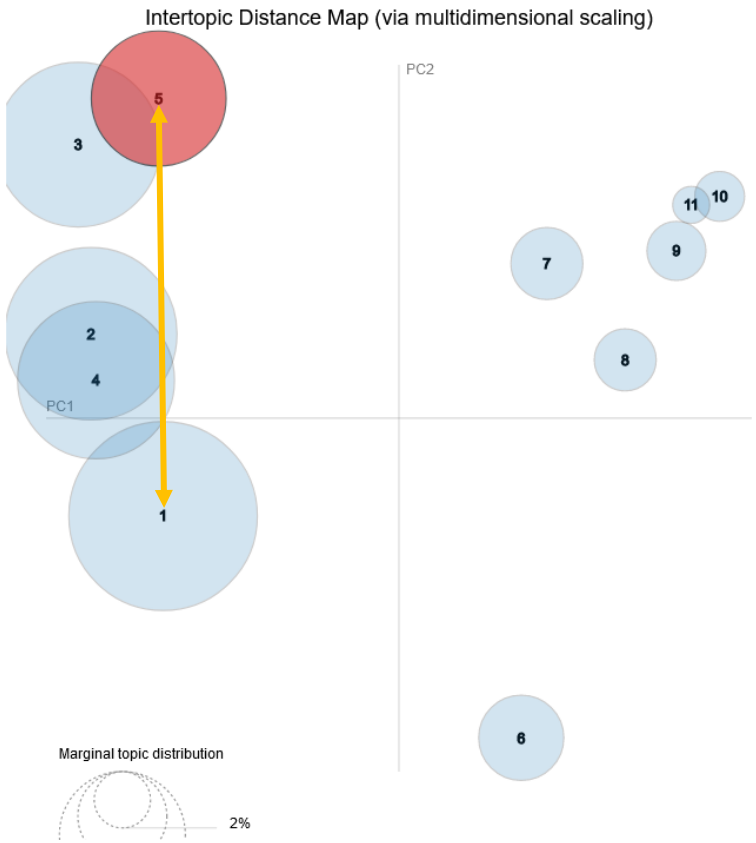


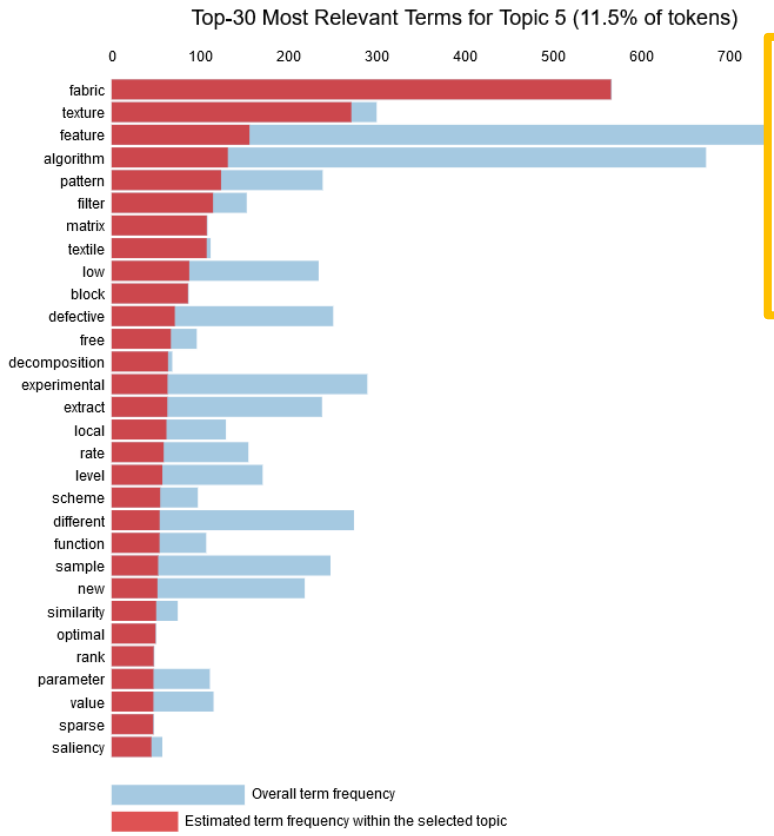
Building LDA model

- Number of Topics;
- Perplexity (inflection);
- Coherence (minimum);
- Hyperparams opt.

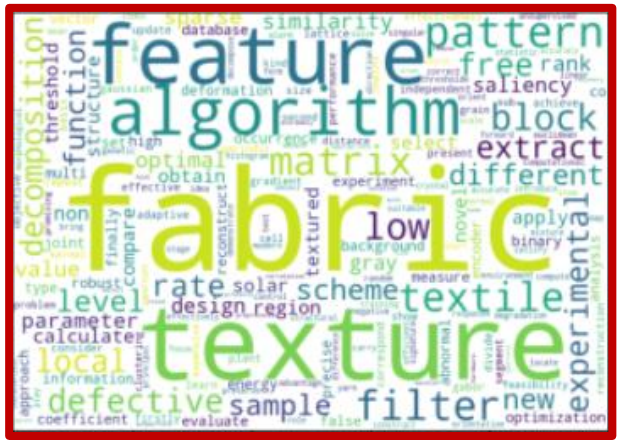


2.3 Topic modeling - *output*





"*Intertopic Distance Plot*" shows topics relate to each other, including potential higher-level structure between groups of topics

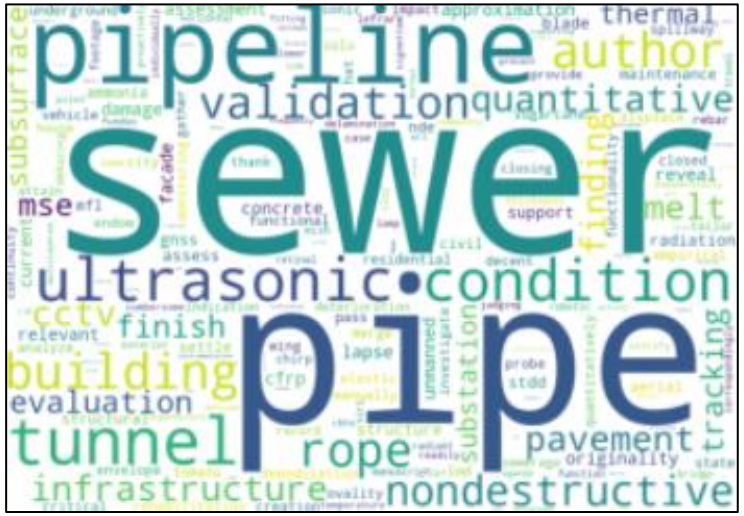


3 Result

The analysis highlighted different topics related to the “automatic defect detection” themes.

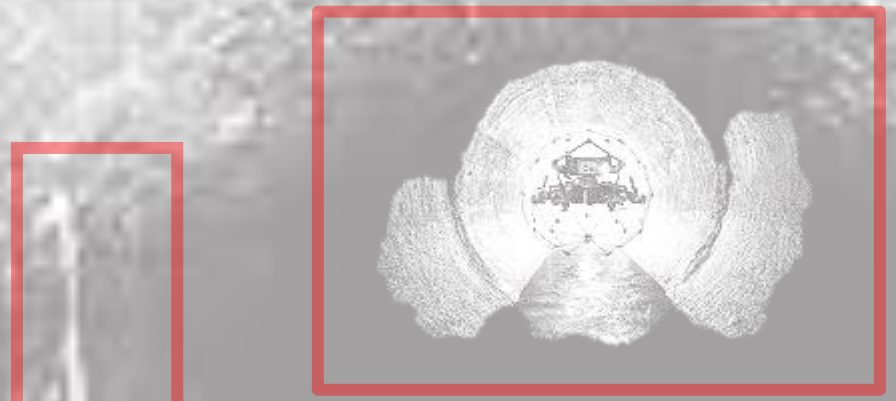
The most of the topics extracted are related to **industry sectors**. One other big part of results are related to **improvement methods** to rise the detections’ checks.

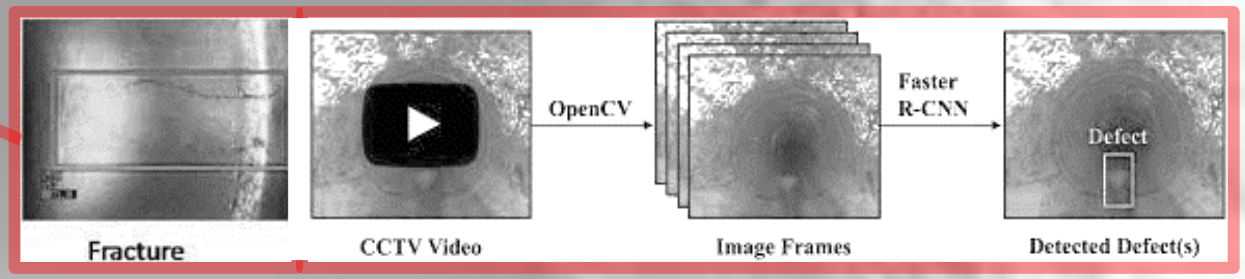
Only two topics are related to the **infrastructure** world, but one of these topics have important words matching specific terms related to the **water utility** industries.



It seems to be some strong applications of the tool for **sewers’** and **pipelines’** field.

Now it’s possible to go deeper to understand which are the techniques used and how much money the company could save apping such results.





Thank you for your attention